Universal Risk Phenotype Of US Counties For Flu-like Transmission To Improve County-specific COVID-19 Incidence Forecasts

Ishanu Chattopadhyay (✉ ishanu@uchicago.edu)
University of Chicago https://orcid.org/0000-0001-8339-8162
Yi Huang
University of Chicago

Biological Sciences - Article

Keywords: Geo-spatial Risk Measure, Transmission Mechanisms, Information-theoretic Function

Posted Date: February 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-124335/v1

Universal Risk Phenotype Of US Counties For Flu-like Transmission To Improve County-specific COVID-19 Incidence Forecasts

Yi Huang[1] and Ishanu Chattopadhyay,[1,2,3,★]

[1]Department of Medicine, University of Chicago, Chicago, IL, USA
[2]Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL, USA
[3]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL, USA

★To whom correspondence should be addressed: e-mail: ishanu@uchicago.edu.

The spread of a communicable disease is a complex spatio-temporal process[1–3] shaped by the specific transmission mechanism, the survivability of the pathogen outside the host under harsh environmental conditions, and access to new viable hosts broadly determined by the local population characteristics, and its compliance to social distancing policies. While the key factors shaping transmission of influenza and COVID-19 are beginning to be broadly understood[4–15], making precise forecasts on case count and mortality is still difficult. Despite a diversity of approaches being used to model the COVID-19 pandemic[16], a single best model is yet to coalesce. In this study we introduce the concept of a universal geo-spatial risk measure, denoted as the Universal Influenza-like Transmission (UnIT) score, to quantify the risk phenotype of US counties facilitating flu-like transmission mechanisms. The UnIT score is computed as a purely information-theoretic function of past incidence data for seasonal flu epidemics, yet emerges as the dominant factor explaining observed county-specific incidence trends over a range of putative demographic and socio-economic factors for the COVID-19 pandemic[17–22]. The predictive ability of the UnIT score is further demonstrated via county-specific weekly case count forecasts which consistently outperform the best models in the current literature[23]. This study demonstrates that knowledge of past epidemics may be used to chart the course of future ones, if transmission mechanisms are broadly similar, despite distinct disease processes and causative pathogens.

WE are in the midst of a global pandemic caused by the novel coronavirus SARS-CoV-2, and reliable prediction of the future local and national case count is crucial for crafting effective intervention policies. Thus the need for tools that chart the likely course of an epidemic in the human population is now more than ever. The spread of a transmissible virus is shaped by diverse interacting factors that are hard-to-model and respond to[3], including the specific transmission mechanism, the survivability of the pathogen outside the host under harsh environmental conditions, and the ease of access to susceptible hosts – determined in part by the density of the local population, its travel habits[3], and compliance to common-sense social distancing policies. Additionally, the prevalence of pre-existing medical conditions in the local population, and its demographic makeup, might modulate susceptibility of specific hosts to the virus, slowing or accelerating the spread of the disease[24,25]. While a broad set of putative factors shaping the spread of communicable viruses such as the seasonal Influenza and COVID-19 are increasingly becoming clear[4–15], making precise granular actionable forecasts of the case counts over time is still difficult. At present, faced with the challenge of forecasting COVID-19 incidence over time, a diversity of modeling approaches have emerged[26–32]. However a single best model is yet to coalesce.

In this study we introduce the concept of a universal geo-spatial risk that strongly influences person-to-person transmission of influenza-like illnesses in the US, in effect quantifying the risk exposure of the individual US counties to diseases that transmit like the seasonal Influenza. We call this the Universal Influenza-like Transmission (UnIT) score. This epidemiologically informed risk score leverages subtle geospatial patterns underlying the seasonal flu epidemics of the past automatically — without the human modelers requiring to manually identify every putative covariate that impacts the process — and significantly improves incidence forecasts for COVID-19 over currently proposed state of the art models. Our results show that the UnIT score emerges as the most important factor “explaining” observed county-specific incidence trends for COVID-19 in the US, with coefficients in multi-variate regression dominating those for typical co-variates. Here our key insight is that past incidence patterns for a different pathogen can substantially inform current projections, provided the transmission mechanisms are broadly similar. We operationalize this insight by crafting a general principle to transfer this

TABLE I
GENERALIZED LINE MODEL RESULT SUMMARY FOR CASE TOTAL AS OF 2020-12-05*

	description	coef.	z-value	0.025	0.975
pop	total population	0.080	1377.188	0.080	0.080
%65+	percentage of population over 65 years old	−0.119	−278.387	−0.120	−0.118
%minority	percentage of minority (non-white) population	−0.029	−35.118	−0.030	−0.027
%black	percentage of black population	−0.000	−0.414	−0.002	0.001
%hispanic	percentage of hispanic population	0.004	12.339	0.003	0.004
%poverty	percentage of population in poverty	−0.166	−243.126	−0.167	−0.164
income	median household income	−0.189	−389.803	−0.190	−0.188
%urban	percentage of urban population	0.074	65.962	0.072	0.076
UnIT	risk phenotype of US counties	0.112	102.536	0.110	0.114
urban-UnIT	UnIT-risk phenotype scaled up by %urban	1.002	720.492	0.999	1.004

*All p-values are < 0.0005.

knowledge from observations of past epidemics to the new context of COVID-19, via a new computable measure of intrinsic similarity between stochastic sample paths generated by the hidden processes.

Intuitively, the ability to leverage Influenza infection patterns to inform COVID-19 modeling is not surprising. COVID-19 and Influenza have similar disease presentations: both are respiratory disorders, which present as a wide range of illnesses from asymptomatic or mild through to severe disease and possible death. Both viruses are transmitted by contact, droplets and fomites[33]. The similarity of the transmission mechanisms is also manifest in the recent drop of reported influenza cases from prevention measures aimed at curbing the spread of COVID-19 worldwide[34–36]. However, to the best of our knowledge, the current paradigms have not capitalized on this dynamical similarity between the transmission mechanisms of the two viruses. The hurdle is that an effective approach to leverage flu patterns in COVID-19 modeling is non-trivial: despite similarities outlined above, there are important empirically observed differences between the two diseases precluding a "drop-in" replacement, *e.g.*, COVID-19 has possibly a higher reproduction number[37–39], can be spread widely by asymptomatic carriers (more so than Influenza[40,41]), is estimated to have a potentially higher mortality rate[42], is novel, *i.e.*, is infecting a host population with almost non-existent immunity, and the COVID-19 pandemic has induced a global trend of social distancing policies alien to the seasonal flu dynamics. Despite these challenges, the UnIT score has significant predictive value, more than manual combinations of putative factors investigated so far.

To put us in context, current COVID-19 modeling efforts may be broadly categorized as: 1) studies that formulate deep theoretical approaches to understand disease propagation in epidemics extending classical compartmental models or their variations[27,43–46]. These investigations generally aim to estimate the theoretical reproduction number of COVID-19, and other epidemiological quantities associated with the virus. 2) While authors in the first category have often striven to connect their models to observational data, the second category of studies aim to forecast incidence, hospitalizations, recovery and deaths more directly, leveraging time-series forecasting tools from statistics and machine learning[28–32,47]. And, 3) in the final category, studies have focused on identifying putative factors driving the differential severity and case counts across regions, demographic strata and age groups[17–22,48–50]. The first category of studies may be seen as theoretical epidemic modeling, the second as predictive analyses (determination of future responses), and the third as inferential analyses[51] (inference of simply how nature associates responses with input variables) to work out the differential impact of putative factors. The current study spans categories 2 and 3, first presenting the UnIT score as a highly explanatory covariate, and then demonstrating its ability to make precise incidence forecasts.

The UnIT risk exposure of a given US county is defined as similarity of the weekly incidence of flu cases to that observed in counties at high risk of epidemic initiation. Thus, central to our approach is the notion of intrinsic similarity between stochastic processes, and how we compute it, particularly if the structure of the underlying processes is unknown. Before we describe the computation of the UnIT score, we need to delineate the concept of sequence likelihood divergence (SLD), which lies at the heart of our computation (See Eq. (8) in Methods). SLD is a rigorous generalization of the notion of divergence of probability distributions (KL divergence[52]) to potentially non-iid stochastic processes. Similar to how we quantify the deviation of a probability distribution p from q by their KL-divergence $\mathcal{D}(p||q)$, SLD allows us to measure the divergence of a stochastic process P from Q as $\mathcal{D}(P||Q)$, where the actual computations are distinct despite the identical notation used (See *Intuitive Example* in Methods). Additionally, the log-likelihood of a sample path x being generated by a process G, denoted as $L(x, G)$, converges in probability with increasing length of x as:

$$L(x, G) \rightarrow H(X) + \mathcal{D}(X||G) \tag{1}$$

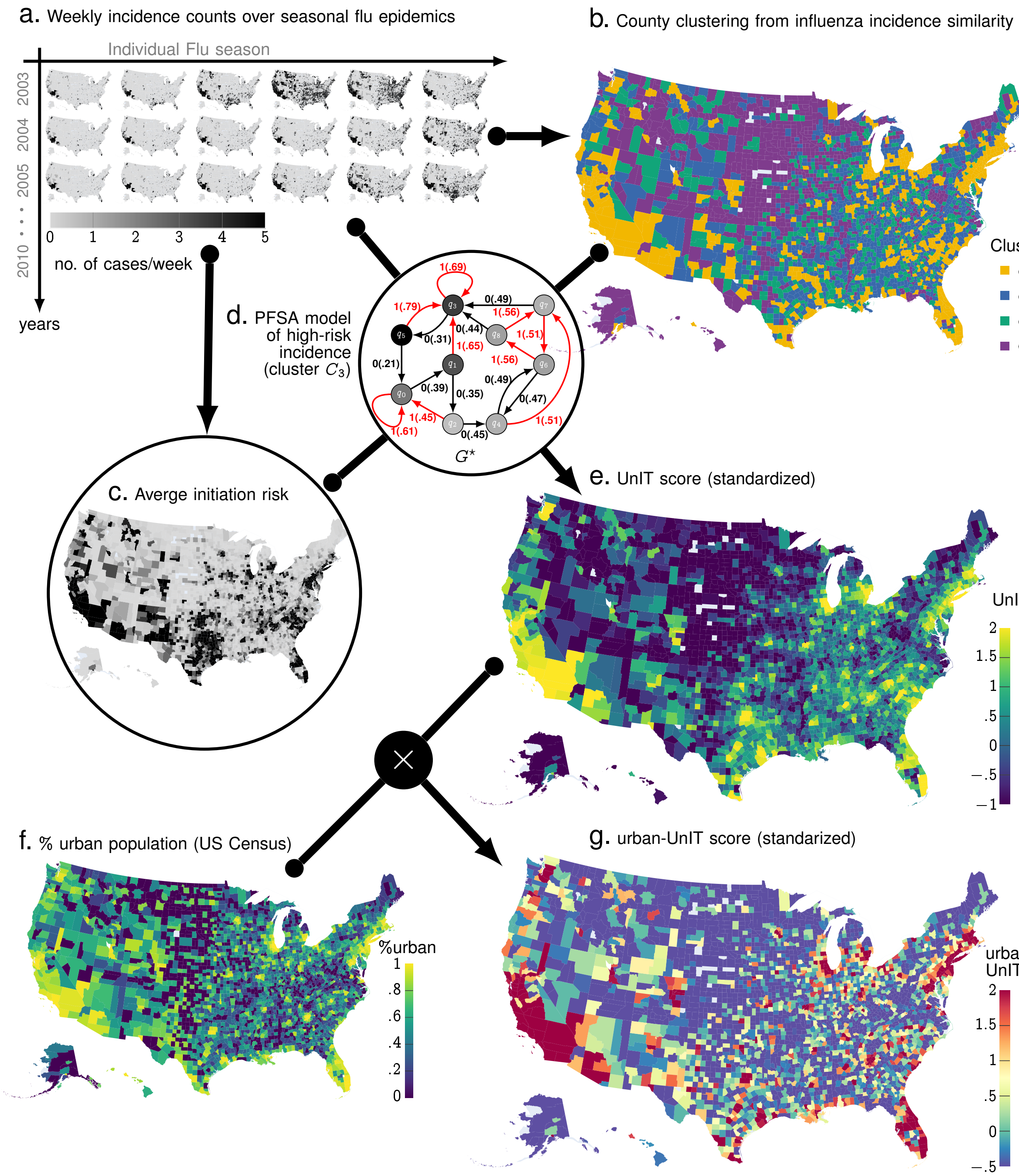


Fig. 1. **Panel a.** Our approach begins with collecting weekly county-wise new case counts of the seasonal flu epidemic spanning Jan. 2003 to Dec. 2012 from a large national database of insurance claims records. We identify weekly Influenza diagnoses using ICD codes related to influenza infection, and end up with county-specific integer-valued time series for each US county for each flu seaon. **Panel b.** These 471-week-long integer-valued time-series are used to compute pairwise similarity between the counties using our new approach of computing intrinsic similarity between stochastic sample paths (See Methods). This similarity matrix induces county clusters C_0, C_1, C_2, C_3. **Panel c.** The flu incidence time series allow us to identify counties which register cases in the first couple of weeks of each flu season. Averaged over all the seasons this gives us a measure of average epidemic initiation risk. **Panel d.** Using the incidence series for the county cluster with maximal average initiation risk we compute a specialized HMM model (PFSA, see Methods) $G^\star$. **Panel e.** Then, we compute the UnIT risk phenotype of each county as the sequence likelihood divergence (SLD, See Methods) between the incience sequence observed and the inferred PFSA model $G^\star$. **Panel f and g**. Finally, the urban-UnIT risk is computed by scaling up the UnIT risk with the fraction of urban population in each county, as obtained from US census (panel f). We show that this risk phenotype is highly predictive of weekly case count of COVID-19, while only dependent of Influenza epidemic history.

where X is the true generator of the sample path x, and $H(\cdot)$ is the entropy rate[52] function (See Methods, Theorem 1). It turns out that if the processes of interest are modeled as a special class of Hidden Markov Models known as Probabilistic Finite State Automata (PFSA)[53], then the estimation of the LHS of Eq. (1)

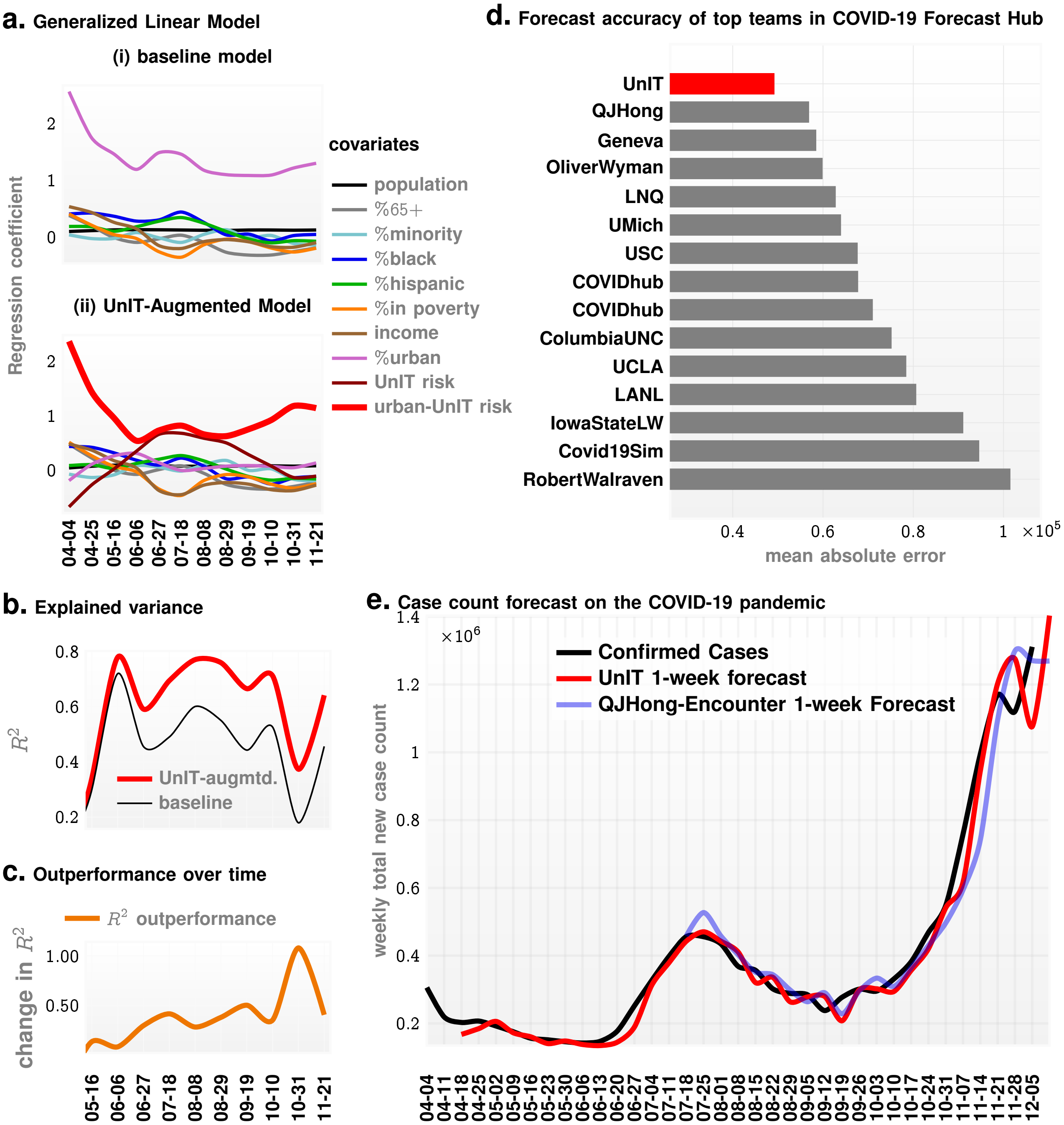


Fig. 2. **Panel a.** We compare the coefficients inferred for the range of covariates shown in the legend for week-wise Poisson regression, for (i) the baseline model and (ii) the model augmented with the UnIT risk variables (See Methods for detailed regression equation). We note that the urban-UnIT risk is significantly dominant for the entire timeline of the pandemic. **Panel b.** The UnIT-augmented has a significantly higher degree of explained variance as measured by R^2. The percentage difference is shown in **panel c**, which demonstrates $> 50\%$ advantage for the major part of the pandemic timeline. **Panel d** illustrates that the UnIT-augmented approach achieves the smallest mean absolute error in one-week ahead county-wise incidence forecasts among the top performing teams from the COVID-19 ForecastHub Community.

becomes tractable (See Methods, Algorithm 1). Hence, using SLD we can efficiently compute the similarity between two observed sample paths, as an intrinsic measure of the deviation between the underlying hidden generators.

Thus, the UnIT risk (denoted as ν) of a county is computed as the SLD between the underlying process driving incidence counts and a high risk process initiating the epidemic. Since these processes are hidden and only sample paths are observable, we formulate an estimator for the UnIT risk as follows: we begin with weekly county-wise confirmed case counts of the seasonal flu epidemic spanning nearly a decade (nine flu seasons between 2003-2012, See Fig. 1a), obtained by looking for Influenza related diagnostic codes in the Truven Marketscan insurance claims database[54]. These relatively short integer-valued time-series (each spanning 471 weeks) are used to compute pairwise similarity between the counties (using the SLD-based approach, see

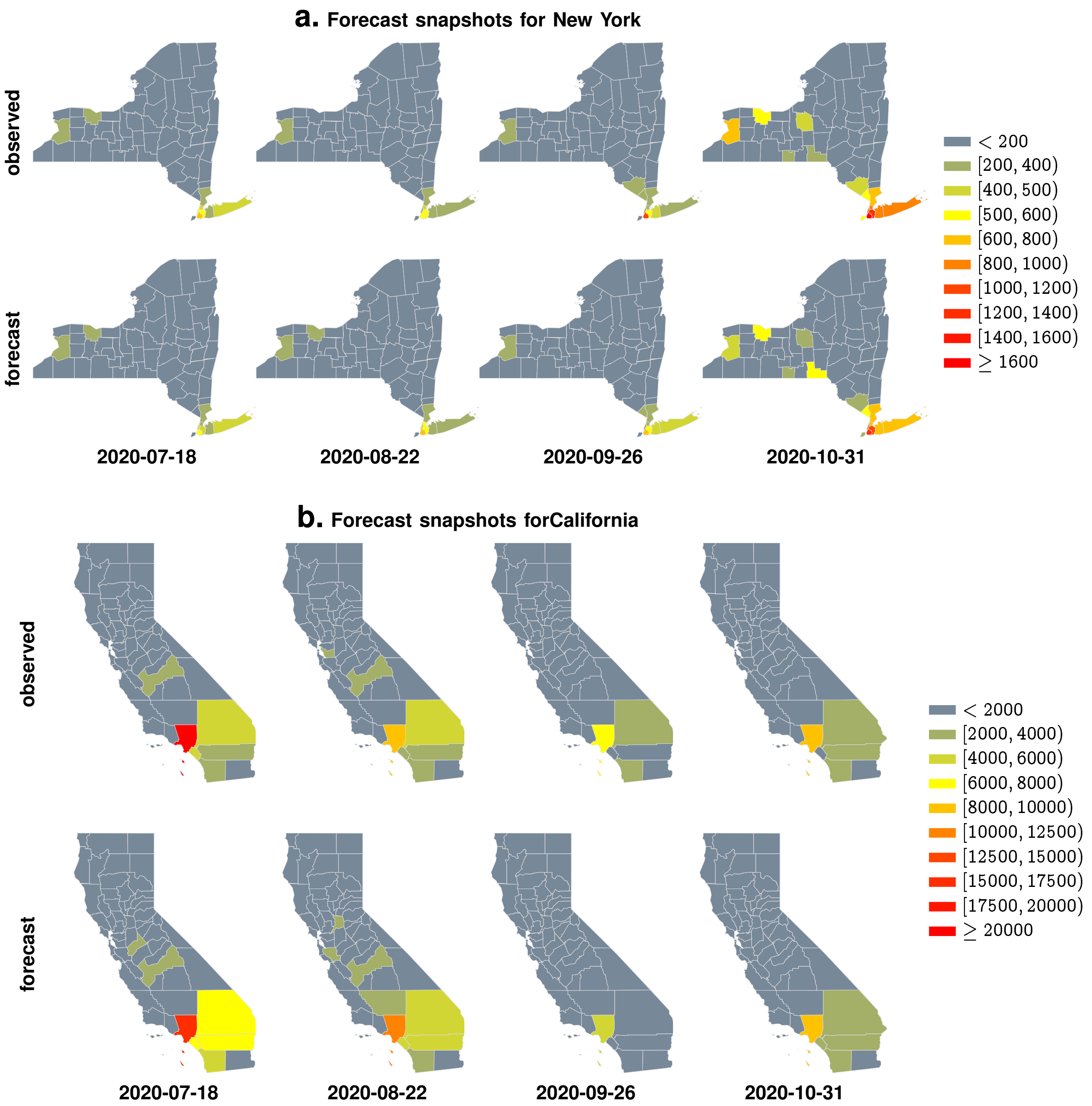


Fig. 3. **Panel a.** We compare our forecasts of weekly case counts (1 week ahead forecasts) with observed confirmed cases on counties from the state of New York. **Panel b.** We compare the weekly forecasts with observed count for the state of California. We note that in both states, for the weeks included in this limited snapshot, the predicted count matches up well with what is ultimately observed.

Methods), which then induces a partition of the 3094 US counties into a pre-specified number of clusters, obtained by using standard clustering techniques, *e.g.* spectral clustering[55] (See Fig. 1b). We note here that the number of clusters (four) is chosen via standard heuristic considerations[56], and increasing this number somewhat does not significantly impact our results. With these county-clusters in hand, we next inspect the initial weeks of the nine flu seasons to estimate the empirical probability of a specific county reporting cases within the first couple of weeks of a flu season — these counties are at high initiation risk empirically (See Fig. 1c). We find that one specific cluster accounts for almost all of the counties at high risk of flu season initiation. Focusing on the set of counties in this high risk cluster, we infer[53] a PFSA $G^\star$, assuming that the incidence series at each of these counties is a sample path from the same underlying stochastic process (See Fig. 1d). This is a simplification, aimed at obtaining an average model driving the incidence dynamics at initiation, ignoring the variation in the structure and parameters of the underlying processes among the high risk counties themselves. Finally, we estimate the UnIT risk exposure of each US county with observed count sequence x as:

$$\widehat{\nu(x)} \triangleq L(x, G^\star) - \widehat{H(X)} \rightarrow \mathcal{D}(X||G^\star) \tag{2}$$

where the convergence to the divergence between the local process X and the inferred high risk process $G^\star$

occurs in probability as length of x increases. To carry out this computation, we need to have a consistent estimate[57] of the entropy rate of the process X from x. This is non-trivial[58,59] if X is not an iid process. We may use one of two approaches here: 1) estimate the entropy rate from the observed sample path[60], or 2) compute an upper bound of the entropy rate assuming X is iid for the purpose of computing $H(X)$ only. The second approach is computationally simpler, but only allows us to estimate a lower bound of the UnIT risk.

To minimize algorithmic complications we present results with only the second approach (See Fig. 1e), $i.e.$ using a lower bound to the UnIT risk, which is demonstrated to have significant predictive value. Additionally, for our COVID-19 modeling, we end up using the estimated urban-UnIT risk in our calculations (See Fig. 1g), which is the product of the UnIT risk and the percentage of urban population in each county. To demonstrate the role of urban-UnIT risk as a meaningful risk phenotype of US counties, we first investigate its role as a co-variate influencing weekly new case count for COVID-19. Diverse putative factors have been investigated with the epidemiological data emerging over the course of the current pandemic, and tested for the strength of their relative contributions influencing the observed incidence variations over time and space. Such variables include weather and pollution covariates[21], population density, socio-economic factors such as poverty, median household income, various measures of income inequality, and fraction of population without medical insurance, demographic variables such as the percentage of African-American, Hispanic and other minorities in the local population, percentage of population aged over 65 years, and gender[17–22]. A common approach here is the use of Poisson regression[61] to establish the statistical significance and relative magnitude of influence of the various factors. We identified the top few variables that have been repeatedly cited as the most important driving factors, and investigated the effect of adding in the urban-UnIT score in multi-variate Poisson regression models, with weekly new case count as the endogenous (target) variable. In our models, the UnIT score dominates with a considerably larger and statistically significant coefficient. Since we standardize all covariates to zero mean and unit standard deviation, the magnitude of the inferred coefficients can meaningfully reflect their relative impact in the models. This is illustrated in Table I where we show the inferred coefficients in a Poisson regression model with the typical covariates along with the urban-UnIT risk. We note that the magnitude of the coefficient for urban-UnIT risk is approximately an order of magnitude larger than that for the covariate with the second largest coefficient magnitude (1.002 for urban-UnIT risk vs -0.189 for median household income). Note that all coefficients inferred are strongly significant with $p < 0.01$. Next we estimate these coefficients for weekly data aggregates, and find that urban-UnIT risk remains dominant over the entire pandemic time-line (See Fig. 2a(ii)). To delineate the predictive power of the new risk variable, we investigate two different models: 1) the baseline model with the covariates outlined in Table I with the exception of the two UnIT risk variables, and 2) the full UnIT augmented model with all the enumerated covariates. The comparative results are shown in panels a(i) and a(ii) of Fig. 2. Comparing the explained variance of the weekly confirmed case counts via the standard R^2 measure (See Fig. 2b-c), we note that the UnIT-augmented model has greater than 50% advantage over the baseline model, maintaining close to $80\% R^2$ for most of the pandemic time-line. Weekly inference of coefficients for 17 weeks between 08-15-2020 to 11-05-2020 is shown in Table II.

To test the robustness of the UnIT score as a key influencing variable, we tested several different modes of perturbation, namely 1) deleting the top 10% of the counties ranked by the highest number of COVID-19 cases per capita, and 2) randomly selecting only 75% of the counties to include in the analysis. Under all such perturbations, the UnIT score retains its position as the dominant factor in our regression models (See Extended Data Fig. 1 in Methods)

With the notion of the UnIT risk formalized, we investigate our ability to forecast the weekly new COVID-19 confirmed case counts across the US counties. The UnIT score (ν) is a spatially varying time-invariant measure. Thus, to forecast the temporal changes in weekly incidence we consider the past week's case count as a feature in training regressors as follows (where X_t is the observed case count at time t, and $\widehat{X}_t$ is the case count forecast made for t at time $t-1$):

$$\text{UnIT risk correction} \qquad X_t^\star = g_t(X_t, \nu, v_1, \cdots, v_m) \tag{3a}$$

$$\text{Regressor training} \qquad X_t = h_t(X_{t-1}^\star) \tag{3b}$$

$$\text{Forecasting estimate} \qquad \widehat{X}_{t+1} = h_t(X_t^\star) \tag{3c}$$

Here g_t is the generalized multivariate regression model (GLM) which carries out the Poisson regression, fitted with X_t as the target variable, and $\nu, v_1, \cdots, v_m$ as exogenous variables, with a logarithmic link function (See Methods for detailed regression equation). ν is the urban-UnIT risk, and the rest of the variables $v_1, \cdots, v_m$ (as described in Table I) are total population, fraction of population over 65 years, fraction of minorities in the population, fraction of Hispanics, fraction of the population reported as African-American or black, fraction of the population designated to be poor, and the median household income. Including the fraction of population living in urban environments as a separate variable does not change results significantly. In Eq. 3b $X_t^\star$ is the estimate of X_t obtained using the inferred coefficients in g_t, and may be viewed as the noise corrected version of the current case count. Finally, we train a standard regressor between the corrected case count and the count observed in the next time step, and use it for forecasting one-week futures (Eq. 3c). The choice of the specific regressor (random forest, gradient boosting, feed-forward neural networks or more complex variants) does not significantly alter our performance. This is an exceedingly simple model compared to the approaches described in the literature, and is essentially a simple ensemble regressor, with input as a UnIT-corrected case count. Nevertheless we outperform the top state of the art models put forward by the COVID-19 modeling community (https://covid19forecasthub.org/community) in mean absolute error in county-specific incidence count estimates (See Fig. 2d). As examples we illustrate the county-wise predicted and confirmed case counts for New York and California at selected weeks over the pandemic, which shows that our 1-week forecasts match up well with the counts ultimately observed (See Fig. 3).

Our results do not imply that Influenza and COVID-19 are similar in their pathobiology, or that any meaningful parallel can be drawn between clinical course of the two diseases. Indeed, one limitation of our approach is its reduced ability to predict COVID-19-related deaths (See Extended Data Fig. 2a). Our death count forecasts are slightly worse than the top few contributors[23] to the COVID-19 forecasthub. We hypothesize that this reduced effectiveness is attributable to the differences between the clinical progression of Influenza and COVID-19: COVID-19 is arguably a more serious disease compared to the seasonal flu, and while historical flu patterns may be leveraged to predict the number of cases, performance suffers when we attempt to extend the same strategy to predict the mortality arising from COVID-19.

The modeling community has responded to the COVID-19 pandemic with a plethora of high performance forecasting approaches to predict case counts, COVID-19-related hospitalizations and deaths (See Extended Data Table I for an incomplete list). These efforts have engaged a diversity of tools and approaches ranging from county-level meta-population estimates to stochastic compartmental models to fitting Gaussian processes to raw data to survival-convolution models to growth rate dynamics to models that take into account human mobility and social distancing policies explicitly. Predictions from these individual contributing groups have also been used to inform an ensemble forecast[16] which may be seen at a live web-based visualization portal at https://viz.covid19forecasthub.org/ (the COVID-19 forecasthub). As a contribution to this community, we report a precise yet simple model for forecasting case counts; one that operates without explicit social distancing and other hard-to-measure parameters, yet supersedes in performance all operating models at the COVID-19 forecasthub, including the ensemble forecast. Our current 1-week forecast may be viewed at the COVID-19 forecasthub webpage (team: UChicagoCHATTOPADHYAY-UnIT), and complete software with installation and usage instructions (See Methods) is publicly available at https://github.com/zeroknowledgediscovery/unitcov.

In this study we demonstrated that one can leverage the knowledge of the incidence fluctuations in one epidemic to inform another with a broadly similar transmission mechanism, despite differences in the epidemiological parameters and the disease processes themselves. This is an important insight particularly in the light of the current pandemic that has claimed hundreds of thousands of lives. Since it is unlikely that this would be the last pandemic we experience, it is crucial to have a tool that can predict case counts early in the pandemic with little prior data available to fit models, by potentially leveraging historical incidence reports pertaining to a different epidemic of the past.

References

[1] Kretzschmar, M. Measurement and modeling: Infectious disease modeling. *Reference Module in Biomedical Sciences* (2016).

[2] Cliff, A. D. & Haggett, P. Statistical modelling of measles and influenza outbreaks. *Statistical methods in medical research* **2**, 43–73 (1993).

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

[4] Chattopadhyay, I., Kiciman, E., Elliott, J. W., Shaman, J. L. & Rzhetsky, A. Conjunction of factors triggering waves of seasonal influenza. *Elife* **7**, e30756 (2018).

[5] Keeling, M. J. & Rohani, P. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters* **5**, 20–29 (2002). URL ⟨GotoISI⟩://WOS:000175032800005.

[6] Viboud, C. *et al.* Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447–51 (2006). URL https://www.ncbi.nlm.nih.gov/pubmed/16574822.

[7] Colizza, V., Barrat, A., Barthelemy, M. & Vespignani, A. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci U S A* **103**, 2015–20 (2006). URL https://www.ncbi.nlm.nih.gov/pubmed/16461461.

[8] Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc Natl Acad Sci U S A* **106**, 21484–9 (2009). URL https://www.ncbi.nlm.nih.gov/pubmed/20018697.

[9] Balcan, D. & Vespignani, A. Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nat Phys* **7**, 581–586 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21799702.

[10] Eggo, R. M., Cauchemez, S. & Ferguson, N. M. Spatial dynamics of the 1918 influenza pandemic in

england, wales and the united states. *J R Soc Interface* **8**, 233–43 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/20573630.

[11] Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–42 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/24337289.

[12] Shaman, J. & Kohn, M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc Natl Acad Sci U S A* **106**, 3243–8 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19204283.

[13] Chowell, G. *et al.* The influence of climatic conditions on the transmission dynamics of the 2009 a/h1n1 influenza pandemic in chile. *Bmc Infectious Diseases* **12** (2012). URL ⟨GotoISI⟩://WOS:000312375400001.

[14] Gog, J. R. *et al.* Spatial transmission of 2009 pandemic influenza in the us. *PLoS Comput Biol* **10**, e1003635 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24921923.

[15] Charu, V. *et al.* Human mobility and the spatial transmission of influenza in the united states. *PLoS Comput Biol* **13**, e1005382 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28187123.

[16] Ray, E. L. *et al.* Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *MedRXiv* (2020).

[17] Zhang, C. H. & Schwartz, G. G. Spatial disparities in coronavirus incidence and mortality in the united states: an ecological analysis as of may 2020. *The Journal of Rural Health* **36**, 433–445 (2020).

[18] Khazanchi, R. *et al.* County-level association of social vulnerability with covid-19 cases and deaths in the usa. *Journal of general internal medicine* **35**, 2784–2787 (2020).

[19] Ehlert, A. The socioeconomic determinants of covid-19: A spatial analysis of german county level data. *medRxiv* (2020).

[20] Mollalo, A., Vahedi, B. & Rivera, K. M. Gis-based spatial modeling of covid-19 incidence rate in the continental united states. *Science of The Total Environment* **728**, 138884 (2020).

[21] Luo, Y., Yan, J. & McClure, S. Distribution of the environmental and socioeconomic risk factors on covid-19 death rate across continental usa: a spatial nonlinear analysis. *Environmental Science and Pollution Research* 1–13 (2020).

[22] Sun, F., Matthews, S. A., Yang, T.-C. & Hu, M.-H. A spatial analysis of covid-19 period prevalence in us counties through june 28, 2020: Where geography matters? *Annals of Epidemiology* .

[23] Reich, N. Viz - covid-19 forecast hub — covid-19. https://viz.covid19forecasthub.org/ (2020). (Accessed on 11/29/2020).

[24] Yanez, N. D., Weiss, N. S., Romand, J.-A. & Treggiari, M. M. Covid-19 mortality risk for older men and women. *BMC Public Health* **20**, 1–7 (2020).

[25] Fang, L., Karakiulakis, G. & Roth, M. Are patients with hypertension and diabetes mellitus at increased risk for covid-19 infection? *The Lancet. Respiratory Medicine* **8**, e21 (2020).

[26] Phelan, A. L., Katz, R. & Gostin, L. O. The novel coronavirus originating in wuhan, china: challenges for global health governance. *Jama* **323**, 709–710 (2020).

[27] Pan, A. *et al.* Association of public health interventions with the epidemiology of the covid-19 outbreak in wuhan, china. jama (2020).

[28] Altieri, N. *et al.* Curating a covid-19 data repository and forecasting county-level death counts in the united states. *arXiv preprint arXiv:2005.07882* (2020).

[29] COVID, I., Murray, C. J. *et al.* Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *MedRxiv* (2020).

[30] Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S. & Ciccozzi, M. Application of the arima model on the covid-2019 epidemic dataset. *Data in brief* 105340 (2020).

[31] Fong, S. J., Li, G., Dey, N., Crespo, R. G. & Herrera-Viedma, E. Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak. *arXiv preprint arXiv:2003.10776* (2020).

[32] Ding, G., Li, X., Shen, Y. & Fan, J. Brief analysis of the arima model on the covid-19 in italy. *medRxiv* (2020).

[33] for Disease Control, C. & Prevention. Assessing risk factors for severe covid-19 illness. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html (2020). (Accessed on 11/05/2020).

[34] Wong, N.-S., Leung, C.-C. & Lee, S.-S. Abrupt subsidence of seasonal influenza after covid-19 outbreak, hong kong, china. *Emerging Infectious Diseases* **26**, 2752 (2020).

[35] Olsen, S. J. *et al.* Decreased influenza activity during the covid-19 pandemicunited states, australia, chile, and south africa, 2020. *Morbidity and Mortality Weekly Report* **69**, 1305 (2020).

[36] Soo, R. J. J., Chiew, C. J., Ma, S., Pung, R. & Lee, V. Decreased influenza incidence under covid-19 control measures, singapore. *Emerging infectious diseases* **26**, 1933 (2020).

[37] Al-Raeei, M. The basic reproduction number of the new coronavirus pandemic with mortality for india, the syrian arab republic, the united states, yemen, china, france, nigeria and russia with different rate of cases. *Clinical epidemiology and global health* (2020).

[38] Billah, M. A., Miah, M. M. & Khan, M. N. Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. *PloS one* **15**, e0242128 (2020).

[39] Dharmaratne, S. *et al.* Estimation of the basic reproduction number (r0) for the novel coronavirus disease in sri lanka. *Virology Journal* **17**, 1–7 (2020).

[40] Oran, D. P. & Topol, E. J. Prevalence of asymptomatic sars-cov-2 infection: A narrative review. *Annals of Internal Medicine* (2020).

[41] Leung, N. H., Xu, C., Ip, D. K. & Cowling, B. J. The fraction of influenza virus infections that are asymptomatic: a systematic review and meta-analysis. *Epidemiology (Cambridge, Mass.)* **26**, 862 (2015).

[42] Brady, P. W. *et al.* Trends in covid-19 risk-adjusted mortality rates. *J Hosp Med.* (2020).

[43] Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B. & Sledge, D. The challenges of modeling and forecasting the spread of covid-19. *Proceedings of the National Academy of Sciences* **117**, 16732–16738 (2020). URL https://www.pnas.org/content/117/29/16732. https://www.pnas.org/content/117/29/16732.full.pdf.

[44] Arenas, A. *et al.* A mathematical model for the spatiotemporal epidemic spreading of covid19. *MedRxiv* (2020).

[45] Li, L. *et al.* Propagation analysis and prediction of the covid-19. *Infectious Disease Modelling* **5**, 282–292 (2020).

[46] Contoyiannis, Y. *et al.* A universal physics-based model describing covid-19 dynamics in europe. *International Journal of Environmental Research and Public Health* **17**, 6525 (2020).

[47] Fong, S. J., Li, G., Dey, N., Crespo, R. G. & Herrera-Viedma, E. Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. *Applied Soft Computing* 106282 (2020).

[48] Briz-Redón, Á. & Serrano-Aroca, Á. A spatio-temporal analysis for exploring the effect of temperature on covid-19 early evolution in spain. *Science of the Total Environment* 138811 (2020).

[49] Kim, S. J. & Bostwick, W. Social vulnerability and racial inequality in covid-19 deaths in chicago. *Health education & behavior* **47** (2020).

[50] Cordes, J. & Castro, M. C. Spatial analysis of covid-19 clusters and contextual factors in new york city. *Spatial and Spatio-temporal Epidemiology* **34**, 100355 (2020).

[51] Donoho, D. 50 years of data science. *Journal of Computational and Graphical Statistics* **26**, 745–766 (2017).

[52] Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley-Interscience, New York, NY, USA, 1991).

[53] Chattopadhyay, I. & Lipson, H. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013).

[54] Hansen, L. The truven health marketscan databases for life sciences researchers. *Truven Health Ananlytics IBM Watson Health* (2017).

[55] Ng, A. Y., Jordan, M. I. & Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, 849–856 (2002).

[56] Alshammari, M. & Takatsuka, M. Approximate spectral clustering with eigenvector selection and self-tuned k. *Pattern Recognition Letters* **122**, 31–37 (2019).

[57] Sober, E. Likelihood and convergence. *Philosophy of Science* **55**, 228–237 (1988).

[58] Schürmann, T. & Grassberger, P. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **6**, 414–427 (1996).

[59] Grassberger, P. Estimating the information content of symbol sequences and efficient codes. *IEEE Transactions on Information Theory* **35**, 669–675 (1989).

[60] Chattopadhyay, I. & Lipson, H. Computing entropy rate of symbol sources & a distribution-free limit theorem. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, 1–6 (IEEE, 2014).

[61] Hedeker, D. & Gibbons, R. *Longitudinal data analysis*. Wiley series in probability and statistics (Wiley-Interscience, Hoboken, N.J., 2006). URL http://www.loc.gov/catdir/enhancements/fy0626/2005058221-d.htmlhttp://www.loc.gov/catdir/enhancements/fy0740/2005058221-b.htmlhttp://www.loc.gov/catdir/enhancements/fy0740/2005058221-t.html.

[62] Lin, J., Keogh, E., Lonardi, S. & Chiu, B. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2–11 (ACM, 2003).

[63] Möller-Levet, C. S., Klawonn, F., Cho, K.-H. & Wolkenhauer, O. Fuzzy clustering of short time-series and unevenly distributed sampling points. In *International Symposium on Intelligent Data Analysis*, 330–340 (Springer, 2003).

[64] Navarro, G. A guided tour to approximate string matching. *ACM computing surveys (CSUR)* **33**, 31–88 (2001).

[65] Chen, L., Özsu, M. T. & Oria, V. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 491–502 (ACM, 2005).

[66] Petitjean, F., Ketterlin, A. & Gançarski, P. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition* **44**, 678–693 (2011).

[67] Helstrom, C. W. *Probability and stochastic processes for engineers* (Macmillan Coll Division, 1991).

[68] Hamming, R. W. Error detecting and error correcting codes. *The Bell system technical journal* **29**, 147–160 (1950).

[69] Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P. & Meester, L. E. *A Modern Introduction to Probability and Statistics: Understanding why and how* (Springer Science & Business Media, 2005).

[70] Crutchfield, J. P. The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena* **75**, 11–54 (1994).

[71] Dupont, P., Denis, F. & Esposito, Y. Links between probabilistic automata and hidden markov models: probability distributions, learning models and induction algorithms. *Pattern recognition* **38**, 1349–1371 (2005).

[72] Chattopadhyay, I. & Lipson, H. Data smashing: uncovering lurking order in data. *Journal of The Royal Society Interface* **11**, 20140826 (2014).

[73] Chattopadhyay, I. Causality networks. *arXiv preprint arXiv:1406.6651* (2014).

[74] Dau, H. A. *et al.* The ucr time series classification archive (2018). https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

[75] Bondy, J. & Murty, U. Graph theory (2008). *Grad. Texts in Math* (2008).

[76] Vidyasagar, M. *Hidden markov processes: Theory and applications to biology*, vol. 44 (Princeton University Press, 2014).

[77] Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).

[78] Greene, W. H. *Econometric analysis* (Pearson Education India, 2003).

[79] Hardy, G. Divergent series, with a preface by je littlewood and a note by ls bosanquet, reprint of the revised (1963) edition. *Éditions Jacques Gabay, Sceaux* (1992).

TABLE II
GENERALIZED LINE MODEL RESULT SUMMARY FOR WEEKLY CASES

		pop	%65+	%minority	%black	%hispanic	%poverty	income	%urban	UnIT	urban UnIT
08-15	z-value	189.4	−15.2	75.2	−27.7	139.3	−67.3	−92.1	−17.4	78.8	91.6
	.025	0.062	−0.045	0.305	−0.117	0.230	−0.275	−0.275	−0.141	0.523	0.775
	.975	0.063	−0.035	0.321	−0.102	0.237	−0.259	−0.264	−0.113	0.550	0.809
	coef.	0.062	−0.040	0.313	−0.109	0.233	−0.267	−0.270	−0.127	0.536	0.792
08-22	z-value	181.4	−50.6	66.3	−30.7	80.2	−62.8	−91.1	−11.3	73.3	79.3
	.025	0.066	−0.155	0.287	−0.138	0.146	−0.282	−0.298	−0.104	0.521	0.723
	.975	0.067	−0.143	0.304	−0.121	0.153	−0.265	−0.285	−0.074	0.549	0.760
	coef.	0.067	−0.149	0.296	−0.129	0.149	−0.273	−0.292	−0.089	0.535	0.741
08-29	z-value	190.2	−83.5	40.1	−36.3	11.9	−17.9	−68.9	9.80	68.4	65.8
	.025	0.073	−0.265	0.174	−0.163	0.019	−0.085	−0.229	0.062	0.490	0.610
	.975	0.075	−0.253	0.191	−0.146	0.027	−0.068	−0.216	0.093	0.519	0.648
	coef.	0.074	−0.259	0.182	−0.155	0.023	−0.077	−0.223	0.078	0.504	0.629
09-05	z-value	196.9	−80.3	26.4	−41.5	−27.4	−2.30	−67.5	27.9	63.7	60.3
	.025	0.079	−0.252	0.114	−0.190	−0.058	−0.018	−0.228	0.205	0.456	0.562
	.975	0.080	−0.240	0.132	−0.173	−0.050	−0.001	−0.215	0.236	0.485	0.600
	coef.	0.079	−0.246	0.123	−0.181	−0.054	−0.010	−0.222	0.221	0.470	0.581
09-12	z-value	171.5	−91.5	13.0	−32.7	−64.0	−0.457	−51.6	29.6	52.6	54.6
	.025	0.080	−0.318	0.057	−0.168	−0.151	−0.011	−0.189	0.237	0.405	0.553
	.975	0.082	−0.305	0.078	−0.149	−0.142	0.007	−0.175	0.270	0.436	0.594
	coef.	0.081	−0.311	0.067	−0.158	−0.147	−0.002	−0.182	0.254	0.420	0.573
09-19	z-value	200.0	−104.9	−0.138	−25.0	−53.6	−26.5	−74.2	11.9	40.8	81.8
	.025	0.084	−0.337	−0.011	−0.131	−0.116	−0.129	−0.260	0.074	0.273	0.738
	.975	0.085	−0.324	0.009	−0.112	−0.108	−0.111	−0.247	0.103	0.301	0.774
	coef.	0.085	−0.331	−0.001	−0.121	−0.112	−0.120	−0.254	0.089	0.287	0.756
09-26	z-value	235.0	−132.5	−6.83	−23.5	−39.7	−49.6	−97.1	−7.32	34.5	100.1
	.025	0.089	−0.416	−0.045	−0.123	−0.082	−0.229	−0.339	−0.064	0.213	0.843
	.975	0.091	−0.404	−0.025	−0.104	−0.074	−0.211	−0.326	−0.037	0.238	0.876
	coef.	0.090	−0.410	−0.035	−0.114	−0.078	−0.220	−0.332	−0.051	0.225	0.859
10-03	z-value	196.5	−109.4	6.70	−51.3	−63.6	−54.0	−100.1	3.47	12.7	113.8
	.025	0.081	−0.332	0.023	−0.248	−0.135	−0.255	−0.347	0.010	0.070	0.957
	.975	0.083	−0.321	0.042	−0.229	−0.127	−0.237	−0.334	0.037	0.095	0.990
	coef.	0.082	−0.326	0.033	−0.239	−0.131	−0.246	−0.340	0.024	0.082	0.974
10-10	z-value	204.7	−122.0	6.17	−57.8	−89.6	−58.8	−109.6	11.7	13.1	115.3
	.025	0.083	−0.349	0.019	−0.261	−0.183	−0.262	−0.362	0.061	0.068	0.908
	.975	0.085	−0.338	0.037	−0.243	−0.175	−0.245	−0.349	0.086	0.091	0.940
	coef.	0.084	−0.343	0.028	−0.252	−0.179	−0.254	−0.355	0.073	0.079	0.924
10-17	z-value	204.8	−110.2	−6.97	−39.4	−82.7	−77.4	−127.0	13.6	10.5	129.8
	.025	0.080	−0.287	−0.042	−0.182	−0.156	−0.327	−0.399	0.068	0.048	0.955
	.975	0.081	−0.277	−0.023	−0.165	−0.149	−0.310	−0.387	0.091	0.070	0.984
	coef.	0.081	−0.282	−0.033	−0.174	−0.153	−0.319	−0.393	0.080	0.059	0.970
10-24	z-value	249.9	−129.9	−17.7	−43.1	−89.1	−86.4	−143.7	12.4	−8.93	160.1
	.025	0.084	−0.306	−0.085	−0.185	−0.150	−0.331	−0.411	0.055	−0.056	1.06
	.975	0.085	−0.297	−0.068	−0.169	−0.143	−0.316	−0.400	0.076	−0.036	1.09
	coef.	0.085	−0.302	−0.077	−0.177	−0.147	−0.323	−0.406	0.065	−0.046	1.08
10-31	z-value	235.0	−138.8	−38.5	−34.6	−96.3	−93.7	−144.3	10.4	−27.4	189.7
	.025	0.077	−0.299	−0.172	−0.147	−0.149	−0.334	−0.376	0.041	−0.140	1.17
	.975	0.078	−0.290	−0.155	−0.131	−0.143	−0.320	−0.366	0.060	−0.121	1.20
	coef.	0.078	−0.295	−0.163	−0.139	−0.146	−0.327	−0.371	0.050	−0.131	1.18
11-07	z-value	299.8	−170.7	−61.4	−3.98	−120.5	−113.8	−153.9	18.0	−33.8	214.7
	.025	0.083	−0.315	−0.242	−0.021	−0.162	−0.345	−0.338	0.066	−0.147	1.14
	.975	0.084	−0.307	−0.227	−0.007	−0.157	−0.333	−0.329	0.082	−0.130	1.16
	coef.	0.084	−0.311	−0.235	−0.014	−0.160	−0.339	−0.334	0.074	−0.139	1.15
11-14	z-value	345.9	−193.5	−72.7	−44.5	−185.1	−112.3	−166.8	27.4	−55.3	262.9
	.025	0.084	−0.309	−0.244	−0.144	−0.223	−0.301	−0.316	0.093	−0.209	1.24
	.975	0.085	−0.302	−0.231	−0.132	−0.218	−0.290	−0.308	0.107	−0.195	1.26
	coef.	0.084	−0.305	−0.238	−0.138	−0.221	−0.295	−0.312	0.100	−0.202	1.25
11-21	z-value	395.5	−154.3	−65.5	−40.2	−148.6	−101.4	−159.3	40.2	−30.4	258.0
	.025	0.084	−0.222	−0.199	−0.118	−0.160	−0.247	−0.273	0.131	−0.111	1.13
	.975	0.085	−0.216	−0.187	−0.107	−0.156	−0.237	−0.267	0.145	−0.097	1.15
	coef.	0.085	−0.219	−0.193	−0.112	−0.158	−0.242	−0.270	0.138	−0.104	1.14
11-28	z-value	398.2	−105.7	−17.6	−97.7	−141.0	−82.0	−146.1	46.4	−2.64	234.7
	.025	0.083	−0.153	−0.053	−0.260	−0.156	−0.204	−0.250	0.161	−0.016	1.07
	.975	0.084	−0.147	−0.042	−0.250	−0.152	−0.194	−0.243	0.175	−0.002	1.09
	coef.	0.084	−0.150	−0.047	−0.255	−0.154	−0.199	−0.246	0.168	−0.009	1.08
12-05	z-value	497.0	−96.8	−33.5	−54.9	−104.6	−87.1	−127.2	39.6	38.0	223.3
	.025	0.090	−0.131	−0.093	−0.142	−0.107	−0.200	−0.200	0.130	0.121	0.961
	.975	0.091	−0.126	−0.083	−0.132	−0.103	−0.191	−0.193	0.144	0.134	0.978
	coef.	0.091	−0.128	−0.088	−0.137	−0.105	−0.196	−0.197	0.137	0.128	0.970

Coefficients with p-value in $[0.01, 0.05)$ are colored blue, and those with p-value ≥ 0.05, red. All other p-values are < 0.01.

Methods

Computing Similarity from Sample Paths

Efficiently contrasting and comparing stochastic processes is the key to analyzing time-dependency in epidemiological patterns, particularly where randomness cannot be ignored. For such learning to occur, we need to define either a measure of deviation or, more generally, a measure of similarity to compare stochastic time series. Examples of such similarity measures from the literature include the classical l_p distances and l_p distances with dimensionality reduction[62], the short time series distance (STS)[63], which takes into account of irregularity in sampling rates, the edit based distances[64] with generalizations to continuous sequences[65], and the dynamic time warping (DTW)[66], which is used extensively in the speech recognition community.

A key challenge in the existing techniques is differentiating complex stochastic processes with subtle variations in their generative structures and parameters. When presented with finite sample paths from non-trivial stochastic processes, the state-of-the-art techniques often focus on their point-wise distance, instead of intrinsic differences in their (potentially hidden) generating processes. Our approach addresses this issue and demonstrably differentiates data streams indistinguishable by state-of-the-art algorithms.

Our intuition follows from a basic result in information theory: if we know the true distribution $\mathbf{p}$ of a random variable, we could construct a code[52] with average description length $h(\mathbf{p})$, where $h(\cdot)$ is the entropy of a distribution. If we used this code to encode a random variable with distribution $\mathbf{q}$, we would need $h(\mathbf{p})+\mathcal{D}(\mathbf{p} \,||\, \mathbf{q})$ bits on average to describe the random variable. Thus, deviation in the distributions show up as an additional contribution from the KL divergence term $\mathcal{D}(\cdot||\cdot)$. Generalizing the notion of KL divergence to processes, we can therefore quantify deviations in process dynamics via an increase in the entropy rate by the corresponding divergence.

Intuitive Example

As a more concrete example of the intuition above, consider the following example with sequences of length n generated by two iid processes $\mathscr{P}_1 = B(.5)$ and $\mathscr{P}_2 = B(.8)$, where $B(p)$ is the Bernoulli process with parameter p[67]. Our objective is to estimate deviations in the binary sample paths generated by these processes. Here we choose iid processes for simplicity, which is *not a restriction in general for our approach*. Let us generate sequences of length n and use E_{ij} to denote the expected Hamming distance[68] between sequences generated by $\mathscr{P}_i$ and $\mathscr{P}_j$. It is easy to show that $E_{11} = E_{12} = E_{21} = 0.5n$, which implies that two sequences both generated by $B(.5)$ are *not* more alike than two sequences where one is generated by $B(.5)$ and the other by $B(.8)$. Using the notation:

$$\begin{aligned} h_1 &= h([.5,.5]) = 1, h_2 = h([.8,.2]) = 0.72, \\ d_{12} &= D_{\text{kl}}([.5,.5] \,||\, [.8,.2]) = 0.32, \\ d_{21} &= D_{\text{kl}}([.8,.2] \,||\, [.5,.5]) = 0.28, \end{aligned}$$

and letting $L(x, B(p))$ denote the log-likelihood of $B(p)$ generating x, we define:

$$\mathbf{v}_x = \left[L(x, B(.5)), L(x, B(.8))\right] \tag{4}$$

Then, by law of large numbers[69], we have:

$$\mathbf{v}_x \to \begin{cases} (h_1, h_1 + d_{12}) = (\mathbf{1.0}, 1.32) & \text{if } x \text{ is generated by } B(.5), \\ (h_2 + d_{21}, h_2) = (1.0, \mathbf{0.72}) & \text{if } x \text{ is generated by } B(.8). \end{cases}$$

which now clearly disambiguates the two processes indistinguishable by their expected Hamming distance, and the correct generator may be identified readily as the one corresponding to the index of the smaller entry in $\mathbf{v}_x$. Our approach generalizes this idea to more complex processes, where we cannot make the iid assumption a priori, thus necessitating the generalization of the notion of KL divergence from probability distributions to stochastic processes.

Log-likelihood of Generating a Sequence $L(x, G)$

In the example above, the generating models are used to evaluate log-likelihoods, which are not directly accessible in our target application. The computation of the log-likelihood $L(x, G)$ of a sequence x generated by a process G, is simple (See Algorithm 1) if we restrict our stochastic processes to those generated by Probabilistic Finite State Automata (PFSA)[70–73]. PFSA are semantically succinct and can model discrete-valued stochastic processes of any finite Markov order, and can approximate arbitrary Hidden Markov Models[71] (HMM). Importantly, PFSA model finite valued processes taking values in a finite pre-specified alphabet. Thus, continuous or integer valued inputs must be quantized, in a manner described later.

In the context of the above discussion, we define dissimilarity Θ between observed sequences x, y as:

$$\Theta(x,y) = \sum_{G^i \in \mathbb{G}} |L(x,G^i) - L(y,G^i)| \tag{5}$$

where $G^i \in \mathbb{G}$ is a set of pre-specified PFSA generators on the same alphabet. And using PFSAs for our base models implies that this measure is easily computatble via multiple applications of Algorithm 1. In our approach, we use the set of four PFSA models shown in Extended Data fig. 3a-d as $\mathbb{G}$. Using a different set of models, which generate processes that are sufficiently pairwise distinct, does not significantly alter our results. These particular "base" models are chosen randomly from all possible PFSAs (See next section) with a maximum of 4 states. For a finite number of base models, Eq. (5) does not technically yield a metric. However, one can approach a metric by increasing the number of models included in the base set. Extended Data Fig. 3e-h illustrates a comparison of this approach of comparing time series with the state of the art Dynamic Time Warp (DTW) algorithm. In particular, our approach is significantly faster yet produces a higher separation ratio (ratio of the mean distance between clusters computed by the two algorithms) for the University of California Riverside (UCR) time-series classification archive[74].

Modeling Details: Probabilistic Finite Automata

Definition 1 (PFSA). *A probabilistic finite-state automaton G is a quadruple $(Q, \Sigma, \delta, \widetilde{\pi})$, where Q is a finite set of states, Σ is a finite alphabet, $\delta : Q \times \Sigma \to Q$ called transition map, and $\widetilde{\pi} : Q \times \Sigma \to [0,1]$ specifies observation probabilities, with $\forall q \in Q, \sum_{\sigma \in \Sigma} \widetilde{\pi}(q,\sigma) = 1$.*

We use lower case Greeks (e.g. σ or τ) for symbols in Σ and lower case Latins (e.g. x or y) to denote sequence of symbols, with the empty sequence denoted by λ. The length of a sequence x is denoted by $|x|$. The set of sequences of length d is denoted by Σ^d.

The directed graph (not necessarily simple with possible loops and multi-edges) with vertices in Q and edges specified by δ is called the graph of the PFSA and, unless stated otherwise, assumed to be strongly connected[75].

Definition 2 (Observation and Transition Matrices). *Given a PFSA $(Q, \Sigma, \delta, \widetilde{\pi})$, the observation matrix $\widetilde{\Pi}_G$ is the $|Q| \times |\Sigma|$ matrix with the (q, σ)-entry given by $\widetilde{\pi}(q,\sigma)$, and the transition matrix Π_G is the $|Q| \times |Q|$ matrix with the (q, q')-entry, written as $\pi(q,q')$, given by $\pi(q,q') = \sum_{\sigma:\delta(q,\sigma)=q'} \widetilde{\pi}(q,\sigma)$.*

Both Π_G and $\widetilde{\Pi}_G$ are stochastic, *i.e.* non-negative with rows of sum 1. Since the graph of a PFSA is strongly connected, there is a unique probability vector $\mathbf{p}_G$ that satisfies $\mathbf{p}_G^T \Pi_G = \mathbf{p}_G^T$[76], and is called the stationary distribution of G.

Definition 3 (Γ-Expression). *δ and $\widetilde{\pi}$ may be encoded by a set of $|Q| \times |Q|$ matrices $\mathbf{\Gamma} = \{\Gamma_\sigma | \sigma \in \Sigma\}$, where*

$$\Gamma_\sigma\big|_{q,q'} = \begin{cases} \widetilde{\pi}(q,\sigma) & \text{if } \delta(q,\sigma) = q', \\ 0 & \text{if otherwise.} \end{cases} \tag{6}$$

We extend the definition of the Γ to $\Sigma^\star$ by $\Gamma_x = \prod_{i=1}^n \Gamma_{\sigma_i}$ for $x = \sigma_1 \dots \sigma_n$ with $\Gamma_\lambda = I$, where I is the identity matrix.

Definition 4 (Sequence-Induced Distributions). *For a PFSA $G = (Q, \Sigma, \delta, \widetilde{\pi})$, the distribution on Q induced by a sequence x is given by $\mathbf{p}_G^T(x) = \left[\!\left[\mathbf{p}_G^T \Gamma_x\right]\!\right]$, where $[\![\mathbf{v}]\!] = \mathbf{v} / \|\mathbf{v}\|_1$.*

Definition 5 (Stochastic process Generated by PFSA). *Let $G = (Q, \Sigma, \delta, \widetilde{\pi})$ be a PFSA, the Σ-valued stochastic process $\{X_t\}_{t \in \Sigma}$ generated by G satisfies that X_1 follows the distribution $\mathbf{p}_G^T \widetilde{\Pi}_G$ and X_{t+1} follows the distribution $\mathbf{p}_G\left(X_1 \cdots X_t\right)^T \widetilde{\Pi}_G$ for $t \in \mathbb{N}$.*

We denote the probability an PFSA G producing a sequence x by $p_G(x)$. We can verify that $p_G(x) = \left\|\mathbf{p}_G^T \Gamma_x\right\|_1$.

Sequence Likelihood Divergence

Definition 6 (Entropy rate and KL divergence). *The entropy rate of a PFSA G is the entropy rate of the stochastic process G generates[77]. Similarly, the KL divergence of a PFSA G' from the PFSA G is the KL divergence of the process generated by the G' from that of G. More precisely, we have the*

$$H(G) = -\lim_{d \to \infty} \frac{1}{d} \sum_{x \in \Sigma^d} p_G(x) \log p_G(x), \tag{7}$$

and the KL divergence

$$D\left(G \,\|\, G'\right) = \lim_{d \to \infty} \frac{1}{d} \sum_{x \in \Sigma^d} p_G(x) \log \frac{p_G(x)}{p_{G'}(x)}, \tag{8}$$

whenever the limits exist.

We also refer to the KL divergence between stochastic processes as the Sequence Likelihood divergence (SLD).

Definition 7 (Log-likelihood)**.** *The log-likelihood*[77] *of a PFSA G generating $x \in \Sigma^d$ is given by*

$$L(x, G) = -\frac{1}{d} \log p_G(x). \tag{9}$$

Algorithm 1 outlines the steps in computing $L(x, G)$. The time complexity of log-likelihood evaluation is $O\left(|x| \times |Q|\right)$ with input length x and $|Q|$ being the size of the PFSA state set.

Theorem 1 (Convergence of Log-likelihood)**.** *Let G and G' be two irreducible PFSA, and let $x \in \Sigma^d$ be a sequence generated by G. Then we have*

$$L(x, H) \rightarrow H(G) + D\left(G \,\|\, G'\right),$$

in probability as $d \rightarrow \infty$.

Proof: See Supplementary text. ■

From distance matrix to similarity matrix

Let D be the pair-wise distance matrix with $d_{ij} = \Theta(s_i, s_j)$, where s_i is the flu time series of county c_i. Then the affinity matrix A for spectral clustering is chosen as $a_{ij} = \exp\left(-d_{ij}^2/2\right)$.

Incidence Data Source for COVID-19 & Demographic, and Socioeconomic Factors

Data on confirmed cases of COVID-19 were compiled and released at the COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19). The John Hopkins COVID-19 data represent data collated by the US Centers for Disease Control & Prevention (CDC) from individual states and local health agencies. Using the John Hopkins COVID-19 data resource, we obtained county-level confirmed new weekly case counts for all weeks upto the current point in time (10-31-2020) for 3094 US counties. We calculated COVID-19 case per capita using the 2019 population estimate provided by the US Census Bureau generated from 2010 US decennial census (https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-detail.html).

We include five demographic independent variables:

1) total population,
2) percent of the total population aged $65+$,
3) percent of Hispanics in the total population,
4) percent of black/African-American in the total population,
5) percent of minority groups in the total population

For socioeconomic factors, we consider:

1) percent of the total population in poverty and
2) median household income,

Which are also obtained from the US Census Bureau, generated based on the 2010 US decennial census.

Incidence Data Source for Seasonal Influenza Epidemics

The source of incidence counts for seasonal flu epidemic is the Truven MarketScan database[54]. This US national database collating data contributed by over 150 insurance carriers and large, self-insuring companies, contains over 4.6 billion inpatient and outpatient service claims, with over six billion diagnostic codes. We processed the Truven database to obtain the reported weekly number of influenza cases over a period of 471 weeks spanning from January 2003 to December 2011, at the spatial resolution of US counties. Standard ICD9 diagnostic codes corresponding to Influenza infection is used to determine the county-specific incidence time series, which are:

1) **487** Influenza
2) **487.0** Influenza with pneumonia
3) **487.1** Influenza with other respiratory manifestations
4) **487.8** Influenza with other manifestations

Discretization of Input Data & Pre-processing Incidence Counts

Integer-valued incidence input is quantized to produce data streams with a finite alphabet, by choosing $k-1$ cut-off points $p_1 < p_2 < \cdots < p_{k-1}$ and replacing a value $< p_1$ by 0, in $[p_i, p_{i+1})$ by i, and $\geq p_{k-1}$ by k. We call the set of cut-off points a *partition*. In our processing of incidence count data for flu epidemics, we obtain a binary partition by first taking a 1-step difference ($i.e.$, transforming a length-n sequence $x_1, x_2, \ldots, x_{n-1}, x_n$ to $x_2 - x_1, x_3 - x_2, \ldots, x_n - x_{n-1}$), and then replacing each positive value in the resulting sequence by 1 and the remaining, 0. The total effect amounts to marking by 1 a week with a rise in case count and by 0, a decrease or an unchanged count.

Calculation of UnIT Risk

We estimate the UnIT risk via the following 6 steps:

1) Compute pairwise similarity between US counties using the metric Θ introduced in Eq. (5).
2) Cluster counties using this similarity measure using standard spectral clustering algorithm[55].
3) Identify the set of counties that have high initiation risk, defined as ones that report cases within the first two weeks of each flu season.
4) Identify the cluster that has a maximal overlap with the set of high-risk counties. If we infer 4 clusters, then we found that only one cluster is sufficient to represent the set of high risk counties. If we set the parameters of the clustering algorithm to find more clusters, then more than one "high-risk" cluster might emerge, which we then collapse and treat as a single set for the next steps.
5) Generate a single PFSA $G^\star$ based on the quantized incidence series from counties in the high-risk cluster cluster, using a reported abductive inference algorithm[53].
6) Finally, estimate UnIT risk as

$$\widehat{\nu(x)} \triangleq L(x, G^\star) - \widehat{H(X)} \rightarrow \mathcal{D}(X || G^\star) \tag{10}$$

The entropy rate is estimated as the entropy of the distribution of 0s and 1s (length 2 probability vector enumerating the fraction of 0s vs 1s), which provides an upper bound to the entropy rate[52]. Thus, our estimate for the UnIT risk actually gives us a lower bound. More detailed computation of the entropy rate only improves results marginally.

Calculation of urban-UnIT Risk

In our modeling and forecasting investigations pertaining to the problem at hand, we use a scaled version of the UnIT risk denoted as the urban-UnIT risk, which is the county-wise product of the UnIT risk with the fraction of the population living in urban environment, as estimated from the 2010 US census.

Poisson Regression Correction To Case Count Forecast

We fit a generalized linear model[61,78] (GLM) with the assumption that the response variable (county specific weekly case counts confirmed for COVID-19) follows a Poisson distribution, and that the logarithm of its expected value can be modeled by a linear combination of unknown parameters.

Specifically, if the response Y, is assumed to be a count that follows a Poisson distribution with mean μ, then:

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k, \tag{11}$$

where $X_1, X_2, \ldots, X_k$ are explanatory variables (covariates). The counts are for all one-week periods between 04-04-2020 to 2020-12-05. This is also known as Poisson regression or a log-linear model.

To investigate the predictive contribution of the UnIT risk, we explore two models:

- *Baseline model* with the following demographic and socio-economic covariates: percentage of urban population, population, percentage of population above 65 years old, percentage of minority population, percentage of black population, percentage of Hispanic population, percentage of population in poverty, median household income;
- *UnIT-Augmented model* which includes the covariates in the baseline model, with the additional urban-UnIT risk factor discussed above.

Note that for the GLM modeling, we use standard score for all covariates and dependent variables with zero mean and unit variance, $i.e.$, assuming the data for a variable is $x_1, \ldots, x_n$ and let $\hat{\mu}$ and $\hat{\sigma}$ be the sample mean and sample standard deviation, respectively, we transform x_i to $(x_i - \hat{\mu})/\hat{\sigma}$, so that a comparison of the magnitudes of the coefficients reflect the relative importance of the significant covariates.

As described in the main text, we use the GLM model to obtain a "corrected" version of the county-specific case count vector, which is subsequently used to train an ensemble regressor to predict case counts 1 week into future. The precise algorithmic steps are enumerated in Algorithm 2. To reduce variance we train a set $\mathcal{R}$ of regressors in the final step, and report the mean. Here $\mathcal{R}$ consists of a random forest model, an extra trees model and a feed-forward neural network model with a single hidden layer implemented through Tensor Flow.

Forecasting COVID-19-related Deaths Over Time

An almost identical approach is used to forecast COVID-19-related deaths, where we use the same covariates as before, but replace the county-specific case count vector with the county-specific record of COVID-19-related deaths. The modified algorithm for forecasting deaths is enumerated in Algorithm 3, where in the step for training regressors, we also use the case count vectors and its corrected version produced by Algorithm 2.

COVID-19 Forecast-hub and Forecast Community

The community of teams providing weekly forecasts in the https://viz.covid19forecasthub.org/ are enumerated at https://covid19forecasthub.org/community. An incomplete list of the teams, with their self-described strategies is given in Extended Data Table I.

Software Usage Instructions

The complete software is available from https://github.com/zeroknowledgediscovery/unitcov. The following steps are required to download, install and execute our model to obtain the next week's case count and COVID-19-related death count estimates. The pre-requisites is to have a Linux operating system, with standard python 3 installation, along with the latest version of jupyter notebook pre-installed. Then, we execute in order:

1) `git clone https://github.com/zeroknowledgediscovery/unitcov.git`
2) `cd unitcov/forecast_pipeline`
3) `jupyter notebook`

Then, in the jupyter notebook environment, execute the following:

1) `pipeline_data_gathering.ipynb`
2) `pipeline_GLM.ipynb`
3) `pipeline_forecast_case.ipynb` and `pipeline_forecast_death.ipynb`

Alternatively, one can run `forecast_for_next_week.ipynb` which is a combination of the steps above.

Data Management

Data sources have been enumerated in details in the relevant sections above (See Incidence Data Source for COVID-19 & Demographic, and Socioeconomic Factors, and Incidence Data Source for Seasonal Influenza Epidemics). With the exception of Truven Marketscan, the rest of these sources are in the public domain. Models generated by the computations are available in a public repository at https://github.com/zeroknowledgediscovery/unitcov, which also includes a complete working copy of the risk computation and forecasting pipeline.

Acknowledgments

This work is funded in part by the Defense Advanced Research Projects Agency (HR00111890043/P00004). The claims made in this study do not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

Algorithm 1: PFSA Log-likelihood

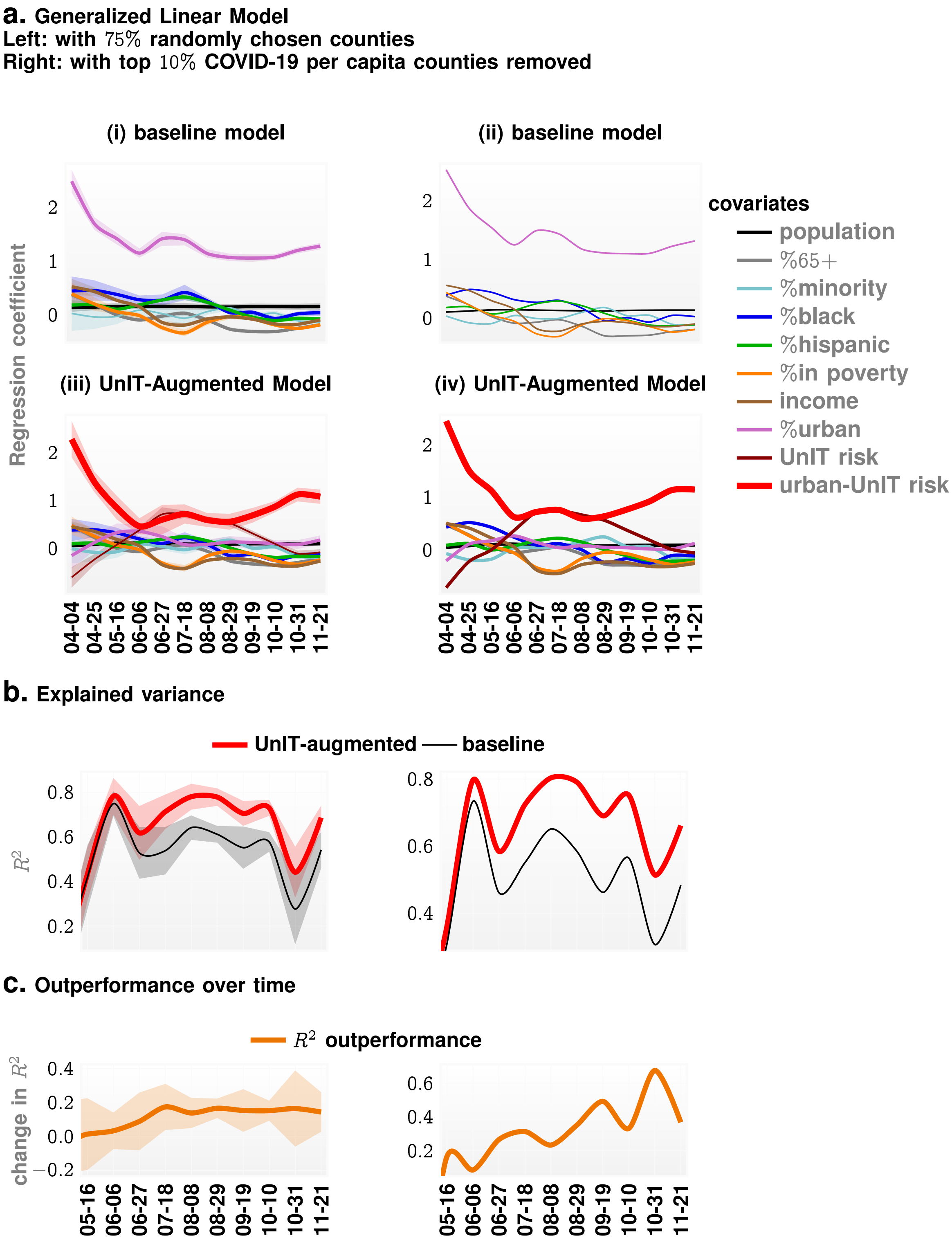
```
Data: A PFSA $G = (Q, \Sigma, \delta, \widetilde{\pi})$ and a sequence $x$ of length $n$.
Result: Log-likelihood of $G$ generating $x$
Get the stationary distribution $\mathbf{p}_G$ as the left eigenvector of $\Pi_G$ of eigenvalue $1$;
Let $\mathbf{p}$ be the current distribution on states, and initialize it with $\mathbf{p}_G$;
Let $L$ be the log-likelihood of $G$ generating $x$ and initialize it with $0$;
for each symbol $\sigma$ in $x$ do
    Get the current distribution on symbols $\phi = \mathbf{p}_G^T \widetilde{\Pi}_G$;
    Update $L = L - \log \phi(\sigma)$;
    Let $\mathbf{p}_{\text{new}}$ be the new distribution on states, and initialize all its entries with $0$;
    for each state $q \in Q$ do
        Let the next the state $q_{\text{new}} = \delta(q, \sigma)$;
        Let $\mathbf{p}_{\text{new}}(q_{\text{new}}) = \mathbf{p}_{\text{new}}(q_{\text{new}}) + \mathbf{p}(q)\widetilde{\pi}(q, \sigma)$;
    Update $\mathbf{p}$ with $\mathbf{p}_{\text{new}} / \|\mathbf{p}_{\text{new}}\|_1$;
Let $L = L/n$;
return $L$;
```

Algorithm 2: Weekly confirmed case forecasting

```
Data:
  • $C_{t-1}$, confirmed cases in time period $t-1$ for each county;
  • $C_t$, confirmed cases in time period $t$ for each county;
  • $C_{glm,t-1}$ and $C_{glm,t}$, approximated case given by the GLM for time period $t-1$ and $t$;
  • A set $\mathcal{R}$ of regressors;
Result: $C_{\text{pred},t+1}$, forecast of confirmed cases in time period $t+1$ for each county.
for each regressor $\texttt{Regr} \in \mathcal{R}$ do
    Let $X_{\text{train}} = [C_{t-1}, C_{glm,t-1}, C_{glm,t}]$;
    Let $y_{\text{train}} = C_t$;
    Fit Regr with $X_{\text{train}}, y_{\text{train}}$;
    Let $X_{\text{pred}} = [C_t, C_{glm,t-1}, C_{glm,t}]$;
    Let $y_{\text{pred},\texttt{Regr}}$ be the prediction Regr makes with $X_{\text{pred}}$;
Let $C_{\text{pred},t+1} = \sum_{r \in \mathcal{R}} y_{\text{pred},r} / |\mathcal{R}|$;
return $C_{\text{pred},t+1}$;
```

Algorithm 3: Weekly death forecasting

```
Data:
  • $D_{t-1}$ and $D_t$, death cases in time period $t-1$ and $t$ for each county;
  • $C_{t-1}$ and $C_t$, confirmed cases in time period $t-1$ and $t$ for each county;
  • $D_{glm,t-1}$ and $D_{glm,t}$, approximated death given by the GLM for time period $t-1$ and $t$;
  • $C_{glm,t-1}$ and $C_{glm,t}$, approximated case given by the GLM for time period $t-1$ and $t$;
  • A set $\mathcal{R}$ of regressors;
Result: $D_{\text{pred},t+1}$, forecast of death in time period $t+1$ for each county.
for each regressor $\texttt{Regr} \in \mathcal{R}$ do
    Let $X_{\text{train}} = [D_{t-1}, C_{t-1}, D_{glm,t-1}, D_{glm,t}, C_{glm,t-1}, C_{glm,t}]$;
    Let $y_{\text{train}} = D_t$;
    Fit Regr with $X_{\text{train}}, y_{\text{train}}$;
    Let $X_{\text{pred}} = [D_t, C_t, D_{glm,t-1}, D_{glm,t}, C_{glm,t-1}, C_{glm,t}]$;
    Let $y_{\text{pred},\texttt{Regr}}$ be the prediction Regr makes with $X_{\text{pred}}$;
Let $D_{\text{pred},t+1} = \sum_{r \in \mathcal{R}} y_{\text{pred},r} / |\mathcal{R}|$;
return $D_{\text{pred},t+1}$;
```



a. Generalized Linear Model
Left: with 75% randomly chosen counties
Right: with top 10% COVID-19 per capita counties removed

b. Explained variance

c. Outperformance over time

Extended Data Fig. 1. To test the robustness of the UnIT score as a key influencing variable, we tested two modes of perturbations: (left column) randomly selecting only 75% of the counties to include in the analysis (considered along with 99% confidence bounds), and (right column) deleting the top 10% of the counties ranked by the highest number of COVID-19 cases per capita. As shown in **panels a** and **b**, under all such perturbations, the UnIT score retains its position as the dominant factor in our regression models, measured by the magnitude of the inferred coefficient relative to those of the other covariates. In particular, in panel a, subpanels (i) and (ii) show the variation of the coefficients for the baseline model for the two perturbation modes described above. The covariates considered in the baseline models are those enumerated in Table I in the main text with the exception of the UnIT risk variables. The corresponding plots for theUnIT-augmented model which includes the additional UnIT risk and urban-UnIT risk as covariates is shown in subpanels (iii) and (iv). **Panel b** shows the explained variation in the models for the two perturbation modes in panels and **panel c** illsurate the outperformance in explained variance.

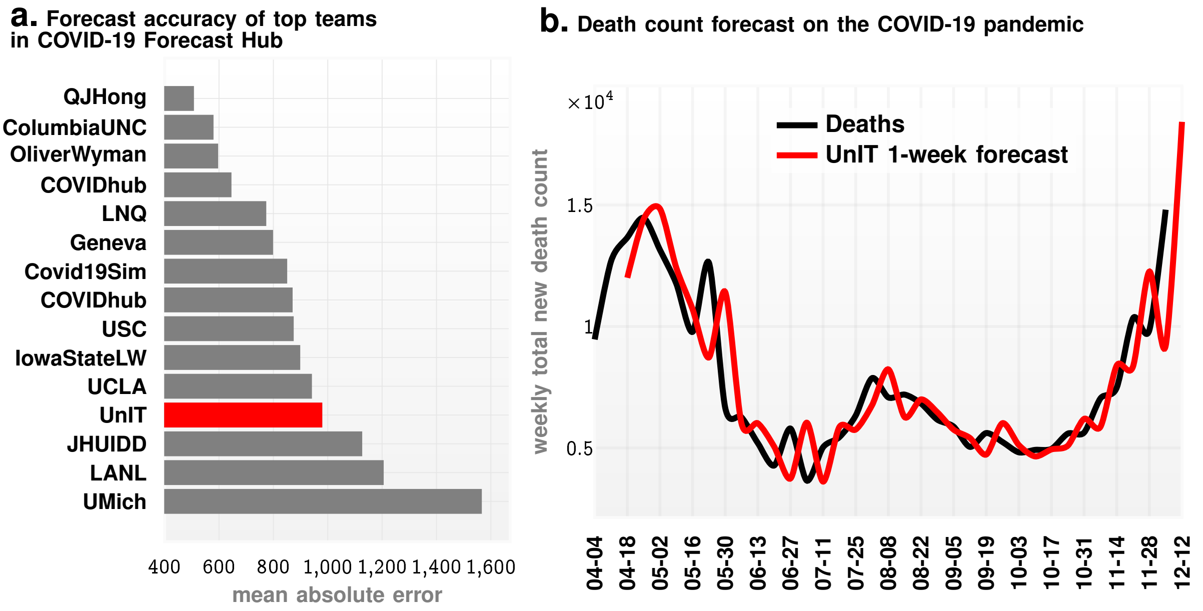
Extended Data Tab. I
COVID-19 FORECASTHUB (HTTPS://COVID19FORECASTHUB.ORG/COMMUNITY) COMMUNITY TEAM SUMMARY

	Team name	Description	Home
1	Johns Hopkins ID Dynamics COVID-19 Working Group (Abbr. JHU_IDD-CovidSP)	County-level metapopulation model with commuting and stochastic SEIR disease dynamics with social-distancing indicators.	https://github.com/HopkinsIDD/COVIDScenarioPipeline
2	Iowa State - Lily Wang's Research Group (Abbr. IowaStateLW-STEM)	Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States.	https://covid19.stat.iastate.edu
3	COVID-19 Simulator (Abbr. Covid19Sim-Simulator)	An interactive tool developed by researchers at Mass General Hospital, Harvard Medical School, Georgia Tech and Boston Medical Center to inform COVID-19 intervention policy decisions in the US.	https://covid19sim.org/
4	Robert Walraven (Abbr. RobertWalraven-ESG)	Multiple skewed gaussian distribution peaks fit to raw data	http://rwalraven.com/COVID19
5	Los Alamos National Labs (Abbr. LANL-GrowthRate)	This model makes predictions about the future, unconditional on particular intervention strategies. Statistical dynamical growth model accounting for population susceptibility.	https://covid-19.bsvgateway.org/
6	UCLA Statistical Machine Learning Lab (Abbr. UCLA-SuEIR)	The SuEIR model is a variant of the SEIR model considering both untested and unreported cases. The model considers reopening and assumes susceptible population will increase after the reopen.	https://covid19.uclaml.org/
7	COVID-19 Forecast Hub (Abbr. COVIDhub-baseline)	This model is a baseline predictive model.	https://covid19forecasthub.org/
8	Columbia_UNC (Abbr. Columbia_UNC-SurvCon)	A survival-convolution model with piece-wise transmission rates that incorporates latent incubation period and provides time-varying effective reproductive number.	https://github.com/COVID19BIOSTAT/covid19_prediction
9	University of Southern California Data Science Lab (Abbr. USC-SI_kJalpha)	A heterogeneous infection rate model with human mobility for epidemic modeling. Our model adapts to changing trends and provide predictions of confirmed cases and deaths.	https://scc-usc.github.io/ReCOVER-COVID-19
10	The University of Michigan (Abbr. UMich-RidgeTfReg)	Nation-level model of confirmed cases and deaths based on ridge regression. No assumptions made about social distancing.	https://gitlab.com/sabcorse/covid-19-collaboration
11	Oliver Wyman (Abbr. OliverWyman-Navigator)	Oliver Wyman's Pandemic Navigator provides forecasts and scenario analysis for Detected and Undetected cases and death counts following a compartmental formulation with non-stationary transition rates.	https://pandemicnavigator.oliverwyman.com/
12	University of Geneva / Swiss Data Science Center (Abbr. Geneva-DetGrowth)	We calculate the growth rate of cumulative cases (resp. deaths) between two days ago and today. If greater than 5%, we use an exponential model to forecast. Otherwise, we use a linear model.	https://renkulab.shinyapps.io/COVID-19-Epidemic-Forecasting/
13	QJHong (Abbr. QJHong-Encounter)	today's "Daily New Confirmed Cases" + today's "Encounter Density" => today's newly infected Cases => next 2-3 weeks' "Daily New Confirmed Cases"	https://qjhong.github.io
14	LockNQuay (Abbr. LNQ-ens1)	County-level ensemble of boosted tree and neural net models. Lots of engineered features.	https://www.kaggle.com/sasrdw/locknquay

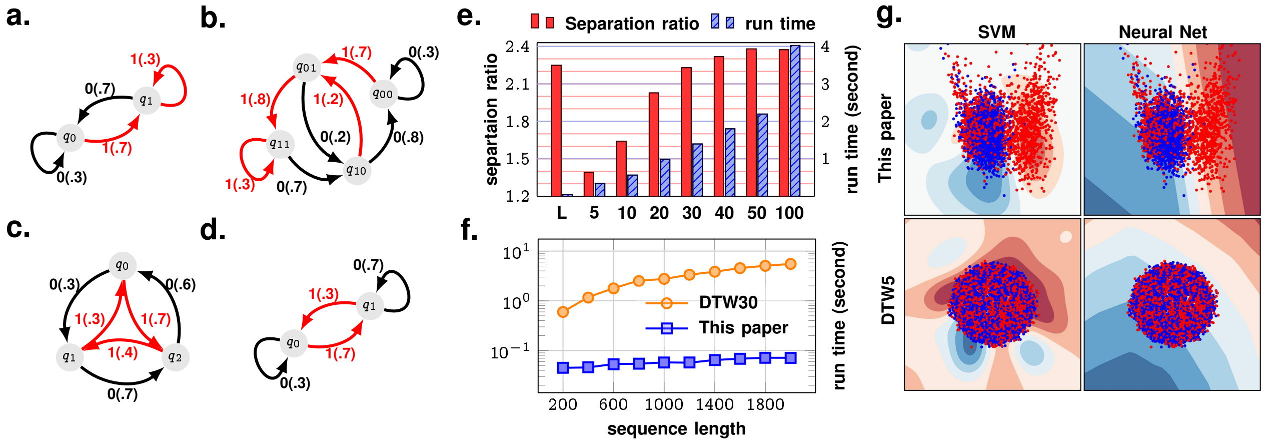
Extended Data Tab. II
GENERALIZED LINE MODEL RESULT SUMMARY FOR WEEKLY DEATH

		pop	%65+	%minority	%black	%hispanic	%poverty	income	%urban	UnIT	urban UnIT
08-15	z-value	29.5	15.0	2.42	3.95	27.9	−8.95	−18.2	−2.29	19.7	10.1
	.025	0.064	0.211	0.018	0.072	0.292	−0.311	−0.494	−0.236	0.877	0.519
	.975	0.073	0.274	0.171	0.213	0.336	−0.199	−0.398	−0.018	1.07	0.769
	coef.	0.069	0.242	0.095	0.143	0.314	−0.255	−0.446	−0.127	0.974	0.644
08-22	z-value	31.9	9.46	2.00	2.41	22.2	−4.05	−14.7	−2.13	17.8	10.2
	.025	0.070	0.130	0.002	0.016	0.235	−0.171	−0.406	−0.224	0.778	0.523
	.975	0.079	0.198	0.153	0.156	0.281	−0.060	−0.311	−0.009	0.970	0.772
	coef.	0.074	0.164	0.077	0.086	0.258	−0.115	−0.358	−0.117	0.874	0.648
08-29	z-value	29.5	12.4	4.05	1.03	20.3	−3.52	−13.9	0.971	18.8	6.57
	.025	0.070	0.181	0.080	−0.033	0.222	−0.163	−0.399	−0.055	0.837	0.301
	.975	0.080	0.249	0.230	0.106	0.270	−0.046	−0.300	0.164	1.03	0.557
	coef.	0.075	0.215	0.155	0.036	0.246	−0.105	−0.350	0.054	0.935	0.429
09-05	z-value	28.4	8.19	8.99	−3.41	19.5	−7.08	−16.8	−2.69	15.7	10.5
	.025	0.066	0.116	0.245	−0.175	0.221	−0.282	−0.484	−0.262	0.692	0.559
	.975	0.076	0.188	0.381	−0.047	0.271	−0.159	−0.383	−0.041	0.889	0.815
	coef.	0.071	0.152	0.313	−0.111	0.246	−0.221	−0.434	−0.151	0.790	0.687
09-12	z-value	25.3	8.90	1.11	2.89	16.3	−1.75	−10.8	−2.67	14.1	10.2
	.025	0.067	0.136	−0.038	0.038	0.194	−0.120	−0.351	−0.271	0.631	0.564
	.975	0.078	0.213	0.136	0.197	0.247	0.007	−0.243	−0.042	0.835	0.832
	coef.	0.073	0.175	0.049	0.117	0.220	−0.057	−0.297	−0.156	0.733	0.698
09-19	z-value	22.3	9.62	6.72	−1.32	15.0	−2.15	−11.0	−0.213	16.5	7.67
	.025	0.060	0.139	0.168	−0.107	0.168	−0.125	−0.320	−0.117	0.694	0.364
	.975	0.071	0.210	0.306	0.021	0.219	−0.006	−0.223	0.094	0.880	0.613
	coef.	0.065	0.175	0.237	−0.043	0.194	−0.065	−0.271	−0.011	0.787	0.488
09-26	z-value	26.4	10.7	3.70	−1.36	11.0	−2.96	−13.2	0.907	11.8	10.2
	.025	0.070	0.157	0.068	−0.121	0.121	−0.160	−0.408	−0.057	0.485	0.535
	.975	0.081	0.228	0.222	0.022	0.173	−0.032	−0.302	0.155	0.679	0.790
	coef.	0.076	0.193	0.145	−0.049	0.147	−0.096	−0.355	0.049	0.582	0.663
10-03	z-value	23.2	5.03	2.75	−2.26	7.42	−1.27	−10.1	0.516	10.0	10.1
	.025	0.067	0.061	0.031	−0.155	0.077	−0.107	−0.325	−0.078	0.398	0.536
	.975	0.079	0.138	0.185	−0.011	0.132	0.023	−0.220	0.134	0.591	0.793
	coef.	0.073	0.099	0.108	−0.083	0.105	−0.042	−0.272	0.028	0.495	0.664
10-10	z-value	22.4	2.88	−0.024	−1.29	4.50	−0.145	−9.40	1.53	9.98	9.64
	.025	0.067	0.018	−0.082	−0.125	0.036	−0.069	−0.303	−0.023	0.387	0.498
	.975	0.080	0.096	0.080	0.026	0.090	0.059	−0.198	0.185	0.577	0.752
	coef.	0.073	0.057	−0.001	−0.050	0.063	−0.005	−0.250	0.081	0.482	0.625
10-17	z-value	20.9	8.82	1.46	−1.64	5.37	−6.59	−13.6	−1.29	7.82	13.8
	.025	0.064	0.123	−0.021	−0.140	0.049	−0.297	−0.432	−0.160	0.262	0.713
	.975	0.077	0.194	0.143	0.012	0.105	−0.161	−0.323	0.033	0.438	0.950
	coef.	0.070	0.158	0.061	−0.064	0.077	−0.229	−0.377	−0.063	0.350	0.832
10-24	z-value	25.4	3.75	4.23	−5.62	−3.02	−7.83	−14.9	5.73	9.97	7.88
	.025	0.078	0.032	0.079	−0.250	−0.073	−0.323	−0.442	0.175	0.341	0.346
	.975	0.091	0.101	0.216	−0.121	−0.015	−0.194	−0.339	0.357	0.508	0.576
	coef.	0.084	0.066	0.148	−0.185	−0.044	−0.258	−0.391	0.266	0.424	0.461
10-31	z-value	22.3	−2.88	3.03	−8.46	−7.18	−6.79	−15.3	2.73	8.90	10.7
	.025	0.071	−0.091	0.037	−0.342	−0.134	−0.285	−0.450	0.035	0.289	0.504
	.975	0.084	−0.017	0.171	−0.213	−0.077	−0.157	−0.348	0.216	0.453	0.730
	coef.	0.077	−0.054	0.104	−0.278	−0.105	−0.221	−0.399	0.126	0.371	0.617
11-07	z-value	32.9	−5.56	−2.47	−1.16	−8.48	−3.56	−13.0	0.980	7.92	13.3
	.025	0.088	−0.132	−0.157	−0.102	−0.137	−0.155	−0.343	−0.040	0.220	0.579
	.975	0.099	−0.063	−0.018	0.026	−0.086	−0.045	−0.253	0.120	0.365	0.778
	coef.	0.093	−0.098	−0.088	−0.038	−0.112	−0.100	−0.298	0.040	0.292	0.679
11-14	z-value	27.7	−4.83	−3.55	−1.87	−8.07	−8.31	−16.3	1.85	6.31	15.5
	.025	0.077	−0.112	−0.203	−0.131	−0.128	−0.295	−0.417	−0.004	0.158	0.674
	.975	0.088	−0.047	−0.059	0.003	−0.078	−0.183	−0.327	0.150	0.300	0.869
	coef.	0.082	−0.080	−0.131	−0.064	−0.103	−0.239	−0.372	0.073	0.229	0.772
11-21	z-value	32.3	−4.71	−5.77	−0.720	−9.73	−9.51	−19.0	5.00	4.16	18.9
	.025	0.078	−0.092	−0.253	−0.081	−0.126	−0.282	−0.409	0.101	0.069	0.725
	.975	0.088	−0.038	−0.125	0.038	−0.084	−0.186	−0.332	0.232	0.193	0.893
	coef.	0.083	−0.065	−0.189	−0.022	−0.105	−0.234	−0.370	0.166	0.131	0.809
11-28	z-value	39.3	−1.67	−4.50	−3.32	−10.7	−10.2	−19.0	3.74	3.90	19.8
	.025	0.089	−0.050	−0.212	−0.162	−0.144	−0.315	−0.420	0.060	0.062	0.774
	.975	0.099	0.004	−0.083	−0.042	−0.100	−0.213	−0.341	0.193	0.188	0.944
	coef.	0.094	−0.023	−0.147	−0.102	−0.122	−0.264	−0.381	0.126	0.125	0.859
12-05	z-value	40.9	−1.39	−4.10	−6.83	−14.0	−10.6	−21.6	7.67	6.90	21.4
	.025	0.080	−0.038	−0.155	−0.213	−0.147	−0.261	−0.374	0.163	0.134	0.714
	.975	0.088	0.006	−0.055	−0.118	−0.111	−0.180	−0.312	0.275	0.240	0.857
	coef.	0.084	−0.016	−0.105	−0.165	−0.129	−0.220	−0.343	0.219	0.187	0.785

Coefficients with p-value in $[0.01, 0.05)$ are colored blue, and those with p-value ≥ 0.05, red. All other p-values are < 0.01.



Extended Data Fig. 2. **Panel a.** Forecast accuracy of COVID-19-related confirmed deaths measured by mean absolute error of top-performaning teams in the COVID-19 forecasthub. **Panel b.** Death count forecasts made by our model against the ground truth. The somewhat reduced effectiveness of our death forecast is probably attributable to the differences between the clinical progression of Influenza and COVID-19.



Extended Data Fig. 3. **Panel a-d.** Four pre-specified PFSA models chosen to act as the basis to estimate similarity between stochastic sample paths. An edge connecting state q to q' is labeled as $\sigma\left(\widetilde{\pi}(q,\sigma)\right)$ if $\delta(q,\sigma) = q'$ (See Defn. 1). **Panel e.** Performance and run time comparisons of SLD distance and DTW on a synthetic dataset. We denote the SLD distance by L and DTW by their window size in Panel e. The average run time of of SLD distance is .042 second. **Panel f.** Run time v.s. sequence length comparison between DTW30 and the SLD distance. Panel g: $2D$ embeddings produced by Alg. 1 and DTW5 on the FordA dataset from the UCR time series classification archive[74] with decision boundaries obtained by using Support Vector Machines (SVM) and neural networks respectively as the classifiers trained with features constructed from the corresponding dissimilarity measures.

Extended Data Tab. III
Generalized line model result summary for death total as of 2020-12-05

	coef.	z-value	.025	.975
pop	0.069	174.380	0.068	0.069
%65+	0.242	95.409	0.237	0.247
%minority	0.120	23.583	0.110	0.130
%black	0.056	12.001	0.047	0.066
%hispanic	0.080	39.387	0.076	0.084
%poverty	0.140	32.421	0.132	0.149
income	0.033	11.232	0.028	0.039
%urban	−0.033	-4.119	-0.049	-0.017
UnIT	0.208	26.708	0.193	0.223
urban UnIT	1.109	113.827	1.090	1.128

All p-values are < 0.0005.

Supplementary Text

Theorem 1 (Convergence of Log-likelihood)**.** *Let G and G' be two irreducible PFSA, and let $x \in \Sigma^d$ be a sequence generated by G. Then we have*

$$L(x, G') \to G'(G) + \mathcal{D}\left(G \,||\, G'\right),$$

in probability as $d \to \infty$.

Proof: By chain rule

$$\begin{aligned}
&\sum_{x\in\Sigma^d} p_G(x)\log\frac{p_G(x)}{p_{G'}(x)}\\
&= \sum_{x\in\Sigma^{d-1}}\sum_{\sigma\in\Sigma} p_G(x)\mathbf{p}_G^T(x)\;\widetilde{\Pi}_G\Big|_\sigma \log\frac{p_G(x)\mathbf{p}_G(x)^T\;\widetilde{\Pi}_G\Big|_\sigma}{p_{G'}(x)\mathbf{p}_{G'}(x)^T\;\widetilde{\Pi}_{G'}\Big|_\sigma}\\
&= \sum_{x\in\Sigma^{d-1}} p_G(x)\log\frac{p_G(x)}{p_{G'}(x)}\\
&\qquad + \underbrace{\sum_{x\in\Sigma^{d-1}} p_G(x)\sum_{\sigma\in\Sigma}\mathbf{p}_G(x)^T\;\widetilde{\Pi}_G\Big|_\sigma \log\frac{\mathbf{p}_G(x)^T\;\widetilde{\Pi}_G\Big|_\sigma}{\mathbf{p}_{G'}(x)^T\;\widetilde{\Pi}_{G'}\Big|_\sigma}}_{D_d}.
\end{aligned}$$

By induction, we have $\mathcal{D}\left(G \,||\, G'\right) = \lim_{d\to\infty}\frac{1}{d}\sum_{i=1}^{d} D_i$, and hence by Cesàro summation theorem[79], we have

$$\mathcal{D}\left(G \,||\, G'\right) = \lim_{d\to\infty} D_d.$$

If $x = \sigma_1\sigma_2...\sigma_n$ is generated by G and $x^{[i-1]}$ is the truncation of x at the $(i-1)$-th symbol, we have

$$-\frac{1}{n}\sum_{i=1}^{n}\log\mathbf{p}_{G'}\left(x^{[i-1]}\right)^T\;\widetilde{\Pi}_{G'}\Big|_{\sigma_i} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathbf{p}_G\left(x^{[i-1]}\right)^T\;\widetilde{\Pi}_G\Big|_{\sigma_i}}{\mathbf{p}_{G'}\left(x^{[i-1]}\right)^T\;\widetilde{\Pi}_{G'}\Big|_{\sigma_i}}}_{A_{x,n}} \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\log\mathbf{p}_G\left(x^{[i-1]}\right)^T\;\widetilde{\Pi}_G\Big|_{\sigma_i}}_{B_{x,n}}.$$

Because the process generated by G is ergodic, we have

$$\lim_{n\to\infty} A_{x,n} = \lim_{d\to\infty} D_d = \mathcal{D}\left(G \,||\, G'\right). \tag{12}$$

and $\lim_{n\to\infty} B_{x,n} = H(G)$. ■

Figures

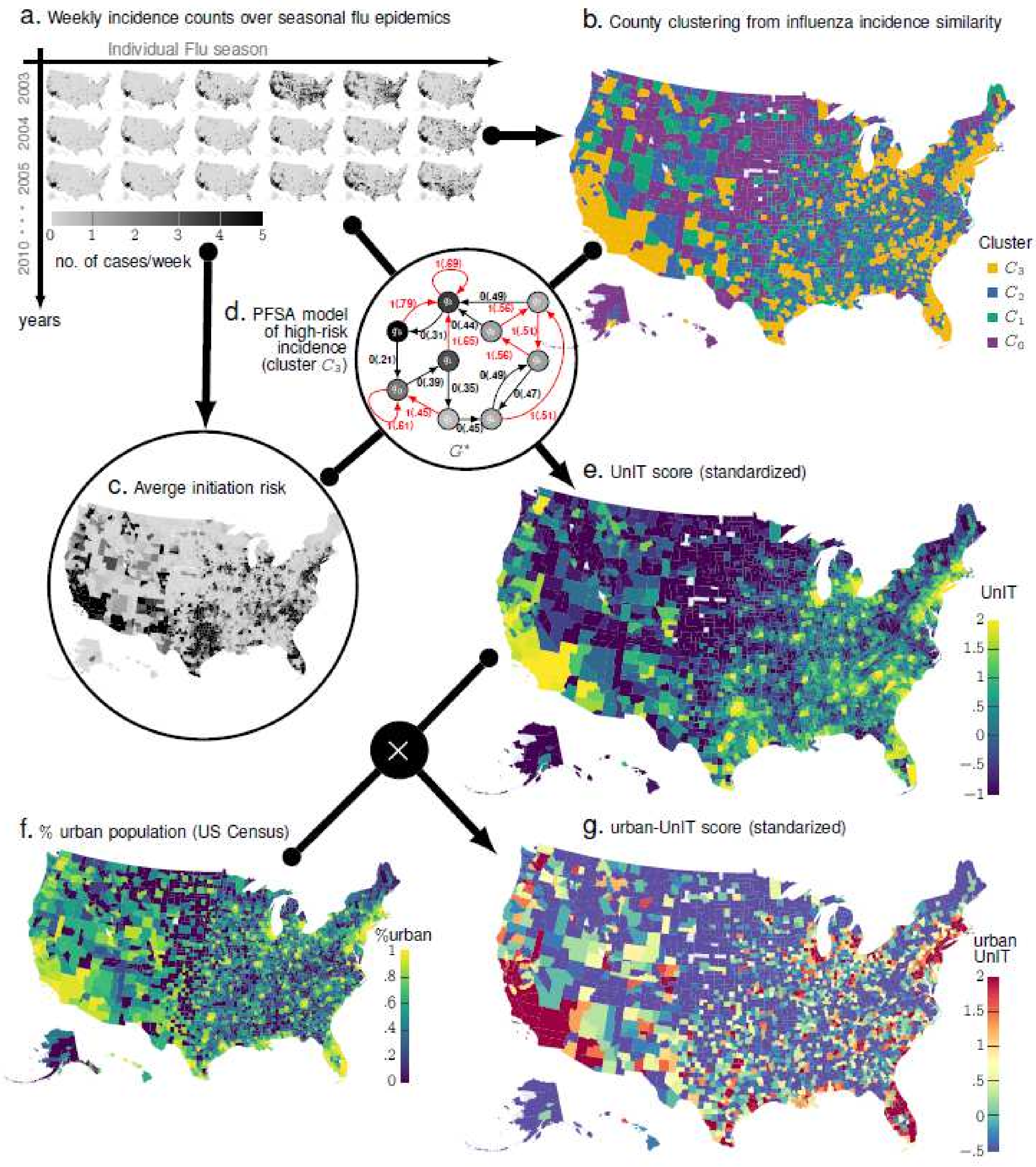

a. Weekly incidence counts over seasonal flu epidemics
Individual Flu season
2003
2004
2005
2010
years
0
1
2
3
4
5
no. of cases/week
b. County clustering from influenza incidence similarity
Cluster
C_3
C_2
C_1
C_0
d. PFSA model of high-risk incidence (cluster C_3)
1(.69)
1(.79)
0(.49)
1(.56)
0(.44)
0(.31)
1(.65)
1(.51)
0(.21)
1(.56)
0(.39)
0(.35)
0(.49)
0(.47)
1(.45)
1(.61)
0(.45)
1(.51)
G^*
c. Averge initiation risk
e. UnIT score (standardized)
UnIT
2
1.5
1
.5
0
−.5
−1
f. % urban population (US Census)
%urban
1
.8
.6
.4
.2
0
g. urban-UnIT score (standarized)
urban UnIT
2
1.5
1
.5
0
−.5

Figure 1

Panel a. Our approach begins with collecting weekly county-wise new case counts of the seasonal flu epidemic spanning Jan. 2003 to Dec. 2012 from a large national database of insurance claims records. We identify weekly Influenza diagnoses using ICD codes related to influenza infection, and end up with county-specific integer-valued time series for each US county for each flu seaon. Panel b. These 471-week-long integer-valued time-series are used to compute pairwise similarity between the counties using our new approach of computing intrinsic similarity between stochastic sample paths (See Methods). This similarity matrix induces county clusters C0;C1;C2;C3. Panel c. The flu incidence time series allow us to identify counties which register cases in the first couple of weeks of each flu season. Averaged over all the seasons this gives us a measure of average epidemic initiation risk. Panel d. Using the incidence series for the county cluster with maximal average initiation risk we compute a specialized HMM model (PFSA, see Methods) G?. Panel e. Then, we compute the UnIT risk phenotype of each county as the sequence likelihood divergence (SLD, See Methods) between the incience sequence observed and the inferred PFSA model G?. Panel f and g. Finally, the urban-UnIT risk is computed by scaling up the UnIT risk with the fraction of urban population in each county, as obtained from US census (panel f). We show that this risk phenotype is highly predictive of weekly case count of COVID-19, while only dependent of Influenza epidemic history.

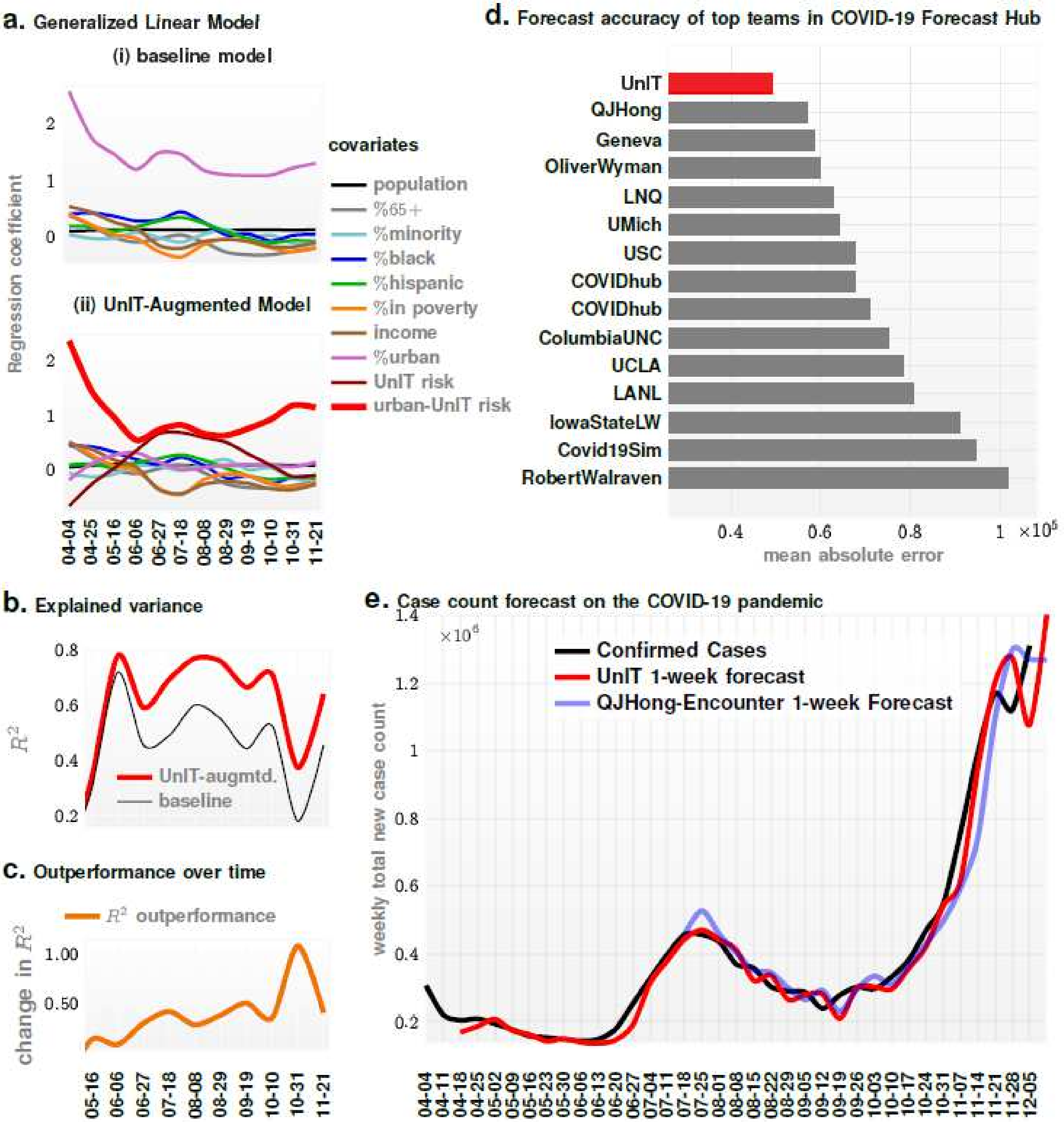


Figure 2

Panel a. We compare the coefficients inferred for the range of covariates shown in the legend for week-wise Poisson regression, for (i) the baseline model and (ii) the model augmented with the UnIT risk variables (See Methods for detailed regression equation). We note that the urban-UnIT risk is significantly dominant for the entire timeline of the pandemic. Panel b. The UnIT-augmented has a significantly higher degree of explained variance as measured by R2. The percentage difference is shown in panel c, which demonstrates > 50% advantage for the major part of the pandemic timeline. Panel d illustrates that the UnIT-augmented approach achieves the smallest mean absolute error in one-week ahead county-wise incidence forecasts among the top performing teams from the COVID-19 ForecastHub Community.

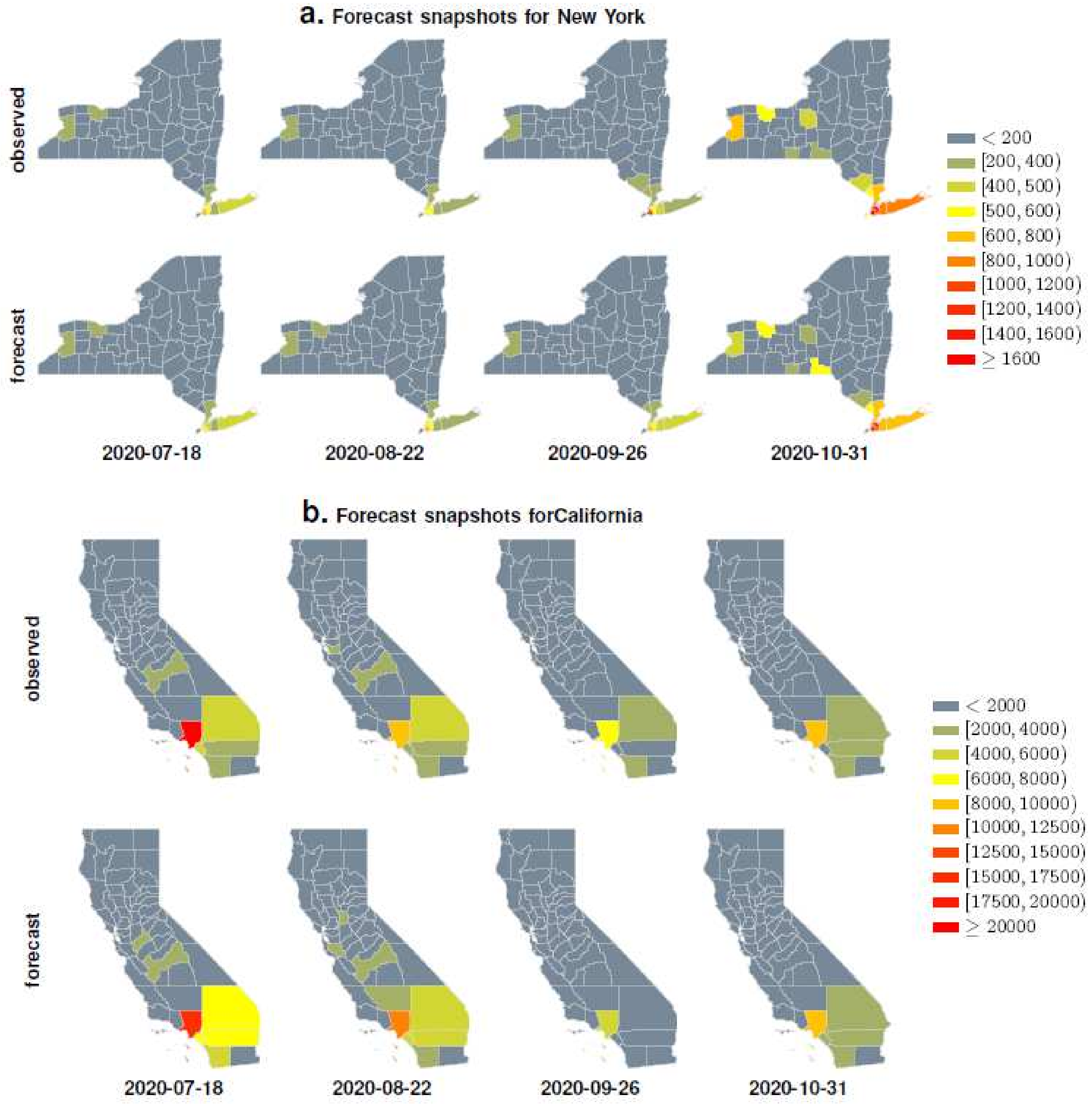


Figure 3

Panel a. We compare our forecasts of weekly case counts (1 week ahead forecasts) with observed confirmed cases on counties from the state of New York. Panel b. We compare the weekly forecasts with observed count for the state of California. We note that in both states, for the weeks included in this limited snapshot, the predicted count matches up well with what is ultimately observed.