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

Yi Huang[a] **and Ishanu Chattopadhyay**[a,b,c]

[a]Department of Medicine, University of Chicago, IL, USA; [b]Committee on Genetics, Genomics & Systems Biology, University of Chicago, IL, USA; [c]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, IL, USA

This manuscript was compiled on Tuesday 5th January, 2021

The spread of a communicable disease is a complex spatio-temporal process shaped by the specific transmission mechanism, and diverse factors including the behavior, socio-economic and demographic properties of the host population. While the key factors shaping transmission of influenza and COVID-19 are beginning to be broadly understood, making precise forecasts on case count and mortality is still difficult. In this study we introduce the concept of a universal geo-spatial risk phenotype of indiidual US counties facilitating flu-like transmission mechanisms. We call this the Universal Influenza-like Transmission (UnIT) score, which is computed as an information-theoretic divergence of the local incidence time series from an high-risk process of epidemic initiation, inferred from almost a decade of flu season incidence data gleaned from the diagnostic history of nearly a third of the US population. Despite being computed from the past seasonal flu incidence records, the UnIT score emerges as the dominant factor explaining incidence trends for the COVID-19 pandemic over putative demographic and socio-economic factors. The predictive ability of the UnIT score is further demonstrated via county-specific weekly case count forecasts which consistently outperform the state of the art models throughout the timeline of the COVID-19 pandemic. 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.

sequence similarity | emergence risk | viral evolution | strain prediction

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 (1), 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 (1), 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 (2, 3). 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 (16–22). However a single best model is yet to coalesce.

Key Insight. In this study we introduce the concept of a universal geo-spatial risk of person-to-person transmission of influenza-like illnesses in the US; universal in the sense that it is pathogen-agnostic provided the transmission mechanism is broadly similar to that of seasonal Influenza. We call this the Universal Influenza-like Transmission (UnIT) score. Transmission dynamics in the general population is known to be modulated by diverse factors, only a few of which have been investigated, and are now beginning to be characterized. In all likelihood many un-modeled factors remain, along with the impact of non-trivial interactions between such known and unknown covriates that are hard to disentangle and account for. The UnIT score allows us to account for the impact of these unmodeled effects by automatically leveraging subtle emergent geospatial patterns underlying the seasonal flu epidemics of the past. In particular, we reduce the need for human modelers to manually identify every putative covariate that impacts the process.

Importantly, the UnIT score – once computed – has applicability beyond seasonal influenza. Validating our claim that the estimated UnIT score indeed quantifies a risk phenotyoe of individual counties for a disease with a flu-like transmission

Significance Statement

A universal risk phenotype of US counties for influenza-like transmission mechanisms is developed which emerges as the dominant variable driving weekly incidence totals amongst the putative set of most widely investigated covariates. Additionally, this risk phenotype allows us to design a simple forecast tool that outperforms the best performing models of incidence prediction reported in the literature. Our results show that knowledge of past epidemics may be effectively used to chart current and future ones, provided the transmission mechanisms are broadly similar, even with distinct pathogens, pathobiologies, and disease presentation.

YH and IC derived the mathematical models, and implemented the algorithm. YH proved the key theorem. IC interpreted results and wrote the paper with significant help from YH.

[2]To whom correspondence should be addressed. E-mail: ishanu@uchicago.edu

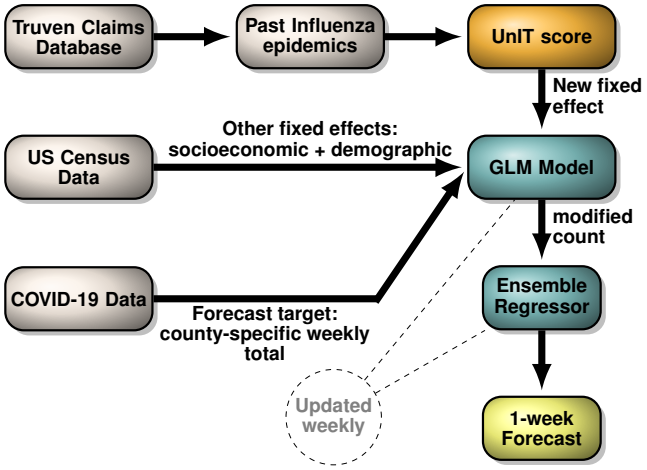


Fig. 1. Modeling Scheme. We use a national insurance claims database with more than 150 million people tracked over a decade (Truven Claims database) to curate geospatial incidence records for past Influenza epidemics over nearly a decade, which informs our new UnIT score. This score is then used as an additional fixed effect along with other putative socio-economic and demographic covariates obtained from US Census to infer a General Linear Model (GLM) explaining the weekly county-specific case COVID-19 case count. Using this inferred GLM model we "correct" the observed weekly case count, and use it as the only feature in an ensemble regressor to forecast county-specific count totals. The GLM model and the regressor a recomputed weekly, while the UnIT score remains invariant, representing a geospatial phenotype modulating transmission.

mechanism, we significantly improve incidence forecasts for COVID-19 over currently proposed state of the art models. We 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 normalized multivariate regression dominating those for typical co-variates. Thus, our key insight is that incidence patterns from a past epidemic caused by a different pathogen can substantially inform current projections under mild assumptions on the similarity of the transmission mechanisms. We operationalize this insight by crafting a general information-theoretic principle to transfer this past knowledge to the new context of COVID-19. This is accomplished via a new computable measure of intrinsic similarity between stochastic sample paths generated by the hidden processes driving incidence.

Modeling Approach. Our overall scheme is summarized in Fig. 1. We leverage the county-specific incidence patterns observed for the past Influenza epidemics to compute the UnIT score, which then is used as a new fixed effect to infer a general linear model (GLM) for county-specific COVID-19 weekly count totals, alongside other putative covariates. The coefficients computed for the GLM model are then used to "correct" the COVID-19 count, replacing the observed count vector with the weighted linear combination of the socio-economic, demographic and the UnIT risk covariates. Intuitively, one may visualize this step as analogous to replacing a somewhat diffused set of observed points with a fitted line in linear regression. Finally, this corrected incidence vector is used to train an ensemble regressor every week that predicts the next weeks county-specific count totals. The GLM model and the regressor is updated weekly, while the UnIT score remains invariant. The key ingredient that makes a simple ensemble regressor in the final step to perform better than more involved tools reported in the literature (See Fig. 3d later) is the information-dense UnIT score, which potentially informs about complex transmission patterns modulating Influenza-like incidence native to each county.

Our ability to leverage Influenza infection patterns to inform COVID-19 modeling is not surprising. COVID-19 and Influenza are both 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 (23). Current efforts to curb the spread of COVID-19 worldwide has also reduced Influenza cases (24–26). However, to the best of our knowledge, the current paradigms have not capitalized on this similarity between the transmission mechanisms of the two viruses. This is not simply an oversight: 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 (27–29), can be spread widely by asymptomatic carriers (more so than Influenza (30, 31)), is estimated to have a potentially higher mortality rate (32), 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.

Results

In our results on COVID-19 modeling, we use a scaled version of the UnIT risk, which we call the urban-UnIT risk (See Fig. 2g). This is the product of the estimated UnIT risk and the percentage of urban population in each county (See Methods for details). To demonstrate the role of urban-UnIT risk as a meaningful risk phenotype of US counties, we first investigate its influence as a covariate driving the weekly total new case count for COVID-19. Over the span of the last few months, diverse putative driving factors have been investigated to explain/model the epidemiological data emerging over the course of the current pandemic. Suspected factors include weather and pollution covariates (33), 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 (33–38). A common approach here is the use of Poisson regression (39) to establish the statistical significance and relative magnitude of the influence of the various individual factors, and their suspected interactions. We identified the 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

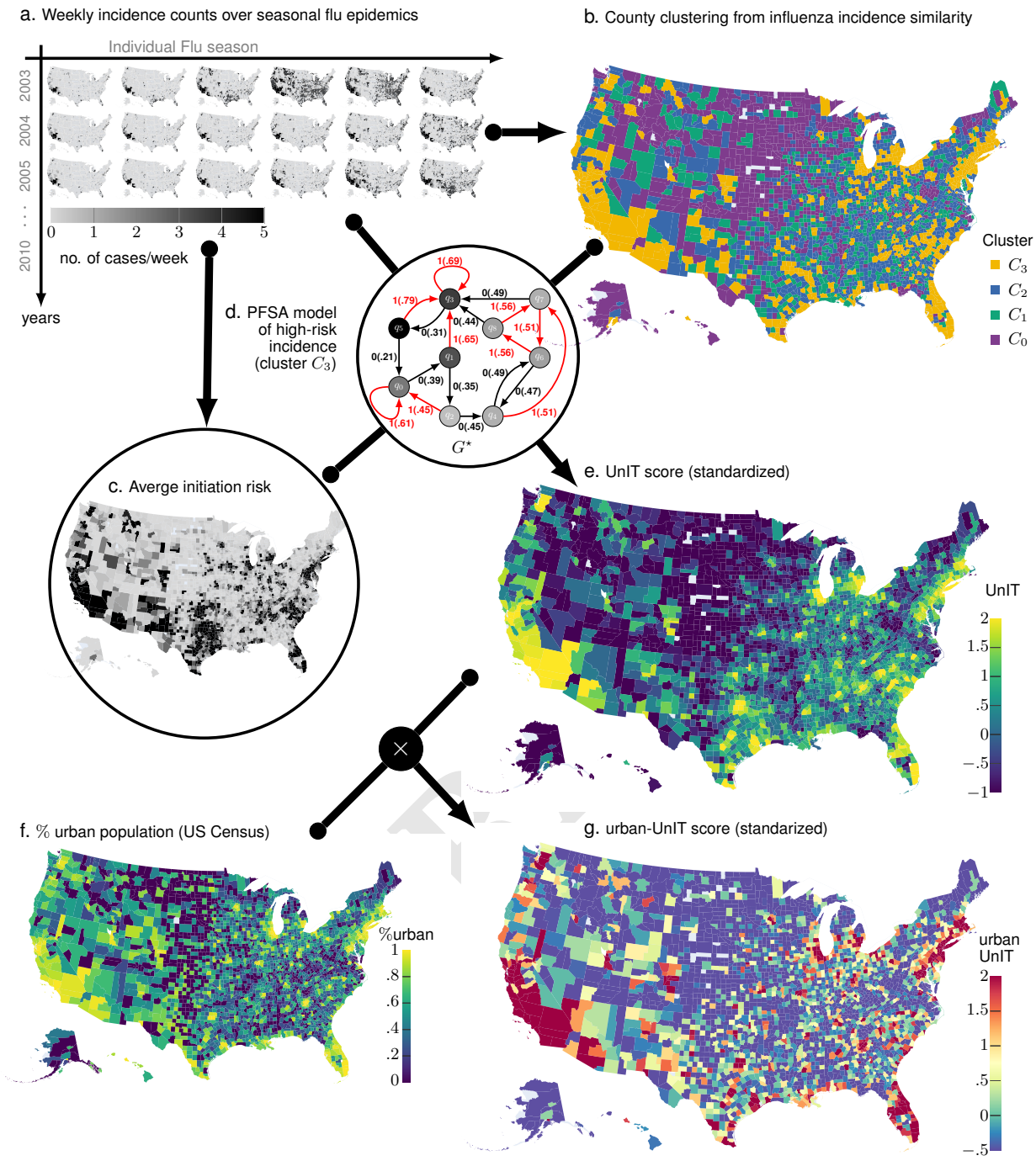


Fig. 2. **Panel a.** Our approach begins with collecting weekly county-wise new case counts of the seasonal flu epidemic spanning Jan. 2003 to Dec. 2012 from a large national database of insurance claims records (Truven Marketscan). We identify weekly Influenza diagnoses using ICD codes related to influenza infection (See Materials and Methods), 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 Eq. (5)). This similarity matrix induces county clusters C_0, C_1, C_2, C_3, inferred via standard spectral clustering. **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 Materials and Methods) $G^\star$. **Panel e.** Then, we compute the UnIT risk phenotype of each county as the sequence likelihood divergence (SLD, See Eq. (8)) 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.

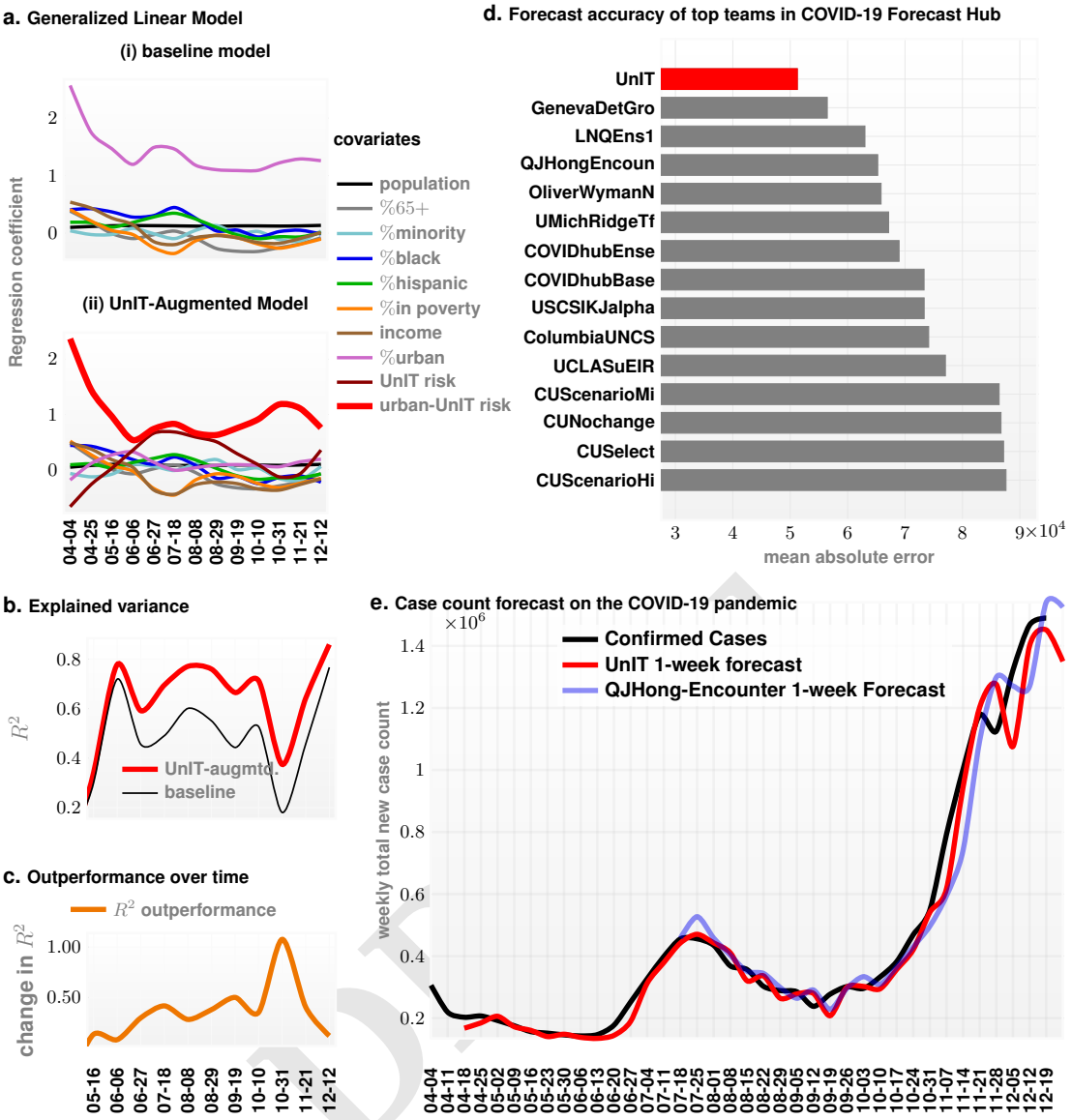


Fig. 3. **Panel a.** We compare the coefficients inferred in multi-variate Poisson regression for individual weeks of the COVID-19 pandemic for the range of covariates shown in the legend. We investigate two models: (i) the baseline model without the UnIT risk related covariates, and (ii) the model augmented with the UnIT risk (See Eq. (11)). We note that the urban-UnIT risk significantly dominates the remaining factors for the entire timeline of the pandemic. **Panel b.** The UnIT-augmented model 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. Finally, **panel e** illustrates the confirmed weekly total of case count summed over all counties vs our 1-week forecast.

case count total as the endogenous (target) variable.

Our first key result is that in our models, the UnIT score significatly dominates typical putative factors. Since we standardize covariates to zero mean and unit variance, the magnitude of the inferred coefficients reflect their relative impact in the models. This is illustrated in Table 1 where we show the inferred coefficients in a Poisson regression model with the typical covariates along with the urban-UnIT risk. In Table 1, we consider county-specific COVID-19 case counts available on 2020-12-19, and note that the magnitude of the coefficient for urban-UnIT risk is approximately an order of magnitude larger than that for the next most influential covariate (0.919 for urban-UnIT risk vs -0.183 for median household income). Note that all coefficients inferred are strongly significant with $p < 0.01$.

Next to demonstrate the dominance of the urban-UnIT risk throughout the current pandemic, we carry out the regression modeling at each week of the current pandemic. We find that urban-UnIT risk remains dominant over the entire pandemic time-line (See Fig. 3a(ii)), by comparing 1) a baseline model with the covariates outlined in Table 1 with the exception of the two UnIT risk variables, vs 2) the full UnIT augmented

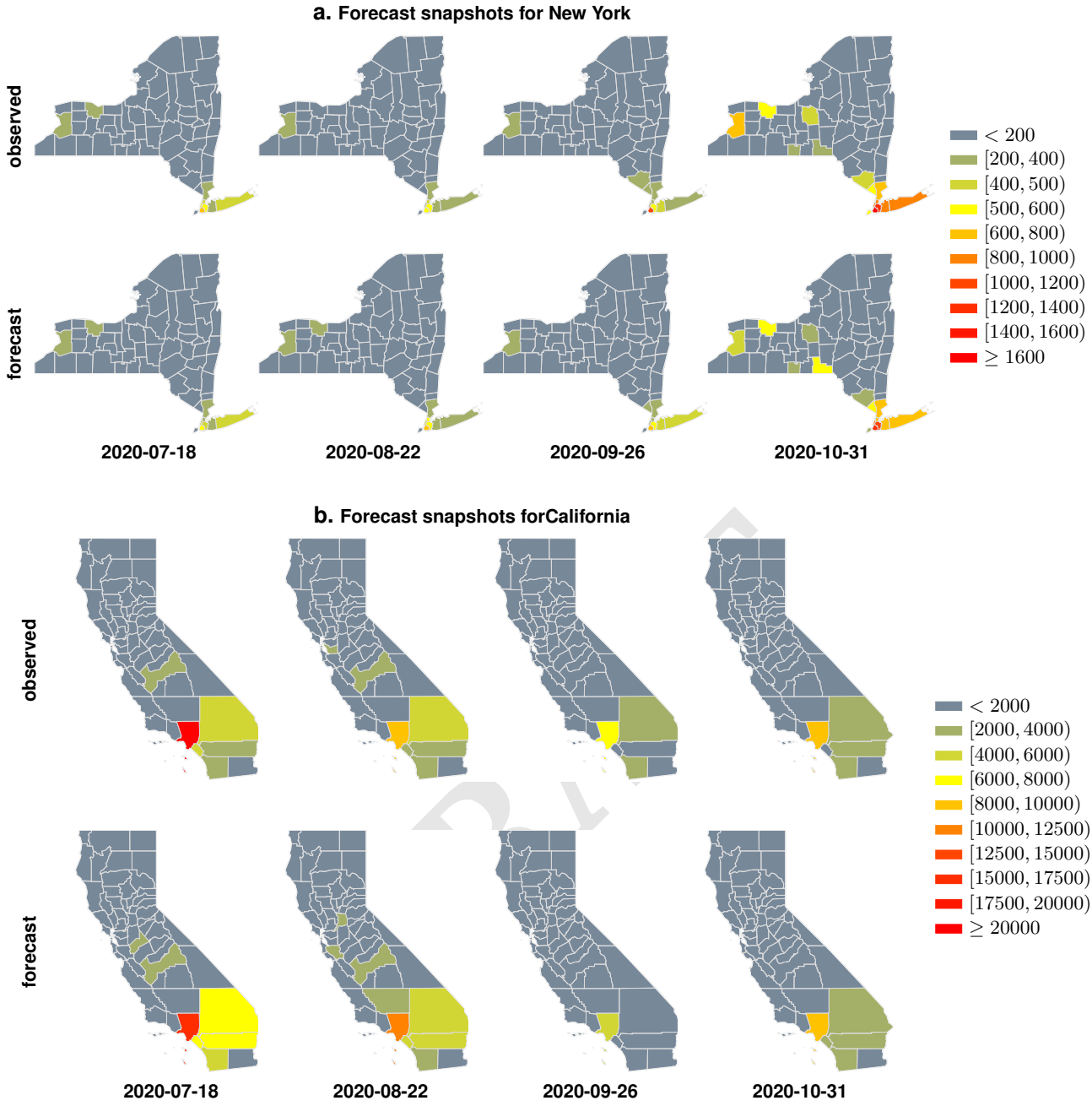


Fig. 4. **Panel a.** We compare our forecasts of weekly case counts (1 week ahead forecasts) with observed confirmed cases on counties from the state of New York. **Panel b.** We compare the weekly forecasts with observed count for the state of California. We note that in both states, for the weeks included in this limited snapshot, the predicted count matches up well with what is ultimately observed.

model with all the enumerated covariates. The comparative results are shown in panels a(i) and a(ii) of Fig. 3. Comparing the explained variance of the weekly confirmed case counts via the standard R^2 measure (See Fig. 3b-c), we note that the UnIT-augmented model has greater than 50% advantage over the baseline model, explaining more than 60% of the variance in the observed weekly COVID-19 case count totals (median $R^2 \approx 0.64$ for augmented model, and ≈ 0.44 for baseline model) for most of the pandemic time-line. Weekly inference of coefficients for past 15 weeks between 2020-09-12 to 2020-12-19 is shown in Table 2.

To demonstrate the robustness of the UnIT score, we investigated multiple modes of perturbation, namely by 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 explanatory factor (See SI Fig. 1 in SI text).

As our second key result, we investigate our ability to forecast weekly COVID-19 case count totals across the US counties. Using a simple forecast model that incorporates the urban-UnIT risk we outperform the state of the

Table 1. Infered coefficients in multi-variate Poisson regression for putative factors driving weekly case totals as of 2020-12-19*

	description	coef.	z-value	0.025	0.975
pop	total population	0.084	1734.008	0.084	0.084
%65+	percentage of population over 65 years old	−0.106	−291.203	−0.107	−0.106
%minority	percentage of minority (non-white) population	−0.003	−4.574	−0.004	−0.002
%black	percentage of black population	−0.056	−86.275	−0.057	−0.054
%hispanic	percentage of hispanic population	−0.006	−21.910	−0.006	−0.005
%poverty	percentage of population in poverty	−0.167	−283.079	−0.168	−0.166
income	median household income	−0.183	−441.725	−0.184	−0.182
%urban	percentage of urban population	0.101	103.694	0.099	0.103
UnIT	risk phenotype of US counties	0.195	207.743	0.193	0.197
urban-UnIT	UnIT-risk phenotype scaled up by %urban	0.919	765.621	0.917	0.921

*All p-values are < 0.0005.

art models from the US COVID-19 modeling community (https://covid19forecasthub.org/community), achieving the least mean absolute error in 1-week ahead county-specific incidence forecasts (See Fig. 3d-e) over the entire pandemic timeline. The predicted and confirmed case counts for New York and California are shown at selected weeks over the pandemic, where our 1-week forecasts match up well with the observed counts (See Fig. 4) in these two US states hit hard by the COVID-19 pandemic.

Discussion

The global modeling community responded to the COVID-19 pandemic with diverse tools (17, 40–43) to predict case counts, COVID-19-related hospitalizations and deaths (See SI Table I for an incomplete list). The proposed approaches range 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. In teh US, predictions from individual contributing groups are been used to inform an ensemble forecast (44), which is currently live at a 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 outperforms the 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 usage instructions (See SI text) is publicly available at https://github.com/zeroknowledgediscovery/unitcov

In addition to the development of forecasting tools, general epidemiological modeling of COVID-19 has progressed in two broad categories: 1) deep theoretical approaches to understand disease propagation in epidemics extending classical compartmental models or their variations (17, 40–43). These investigations aim to estimate the theoretical reproduction number of COVID-19, and other epidemiological quantities associated with the virus. And, 2) in the second category, studies have focused on identifying putative factors driving the differential severity and case counts across regions, demographic strata and age groups (33–38, 45–47). The first category of studies may be seen as theoretical epidemic modeling, and the second as inferential analyses (48), *i.e.*, infer how nature associates responses with input variables aiming to work out the differential impact of putative factors. The current study improves results in the second category by presenting the UnIT score as a highly explanatory covariate, and then demonstrates its ability to make precise incidence forecasts.

The UnIT risk exposure of a US county is conceived of as intrinsic similarity of the time series of weekly total of new flu cases to that observed in counties at high risk of an epidemic initiation. Thus, central to our approach is the notion of intrinsic similarity between stochastic processes, particularly if the structure of the underlying processes is unknown. Such (dis)similarity is quantified by the notion of sequence likelihood divergence (SLD), which lies at the heart of our computation (See Eq. (8) in Materials and Methods). SLD is a generalization of the notion of divergence of probability distributions (KL divergence (49)) 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 measures the divergence of a stochastic process P from Q as $\mathcal{D}(P||Q)$. The actual computations are distinct despite the identical notation used (See *Intuitive Example* in Materials and Methods). Additionally, the log-likelihood of a sample path x being generated by a process G, denoted as $L(x, G)$, converges in probability:

$$L(x, G) \to H(X) + \mathcal{D}(X||G) \qquad [1]$$

with increasing length of x, where X is the true generator of the sample path x, and $H(\cdot)$ is the entropy rate (49) function (See Materials and Methods, Theorem 1). Importantly, if the processes are modeled by a special class of Hidden Markov Models known as Probabilistic Finite State Automata (PFSA) (50), then the estimation of the LHS of Eq. (1) becomes tractable (See SI text, Algorithm 1). Using SLD we can efficiently compute the dissimilarity between two observed sample paths, estimated as the deviation between the underlying generators.

Thus, the UnIT risk (denoted as ν) of a county is defined 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

Table 2. Inferred coefficients in multi-variate Poisson regression for individual weeks

		pop	%65+	%minority	%black	%hispanic	%poverty	income	%urban	pre-UnIT	UnIT
09-12	*z*-value	171.3	−91.2	13.2	−32.8	−63.7	−0.915	−51.8	29.6	52.7	54.6
	.025	0.080	−0.317	0.058	−0.168	−0.150	−0.014	−0.189	0.236	0.405	0.553
	.975	0.082	−0.303	0.079	−0.149	−0.141	0.005	−0.176	0.270	0.437	0.594
	coef.	0.081	−0.310	0.069	−0.159	−0.146	−0.004	−0.182	0.253	0.421	0.573
09-19	*z*-value	200.2	−104.9	−0.104	−25.1	−53.5	−26.5	−74.2	12.3	40.7	81.6
	.025	0.084	−0.336	−0.011	−0.131	−0.116	−0.129	−0.260	0.078	0.272	0.736
	.975	0.086	−0.324	0.010	−0.112	−0.108	−0.111	−0.247	0.107	0.300	0.772
	coef.	0.085	−0.330	−0.001	−0.121	−0.112	−0.120	−0.254	0.092	0.286	0.754
09-26	*z*-value	235.0	−132.5	−6.51	−23.9	−39.6	−49.8	−97.2	−7.09	34.6	100.0
	.025	0.089	−0.416	−0.043	−0.125	−0.082	−0.230	−0.339	−0.063	0.213	0.841
	.975	0.091	−0.403	−0.023	−0.106	−0.074	−0.212	−0.326	−0.036	0.239	0.875
	coef.	0.090	−0.409	−0.033	−0.115	−0.078	−0.221	−0.333	−0.049	0.226	0.858
10-03	*z*-value	196.6	−109.4	6.85	−51.6	−63.7	−53.9	−100.0	3.69	12.6	113.8
	.025	0.081	−0.332	0.024	−0.249	−0.135	−0.255	−0.346	0.012	0.069	0.956
	.975	0.083	−0.320	0.043	−0.231	−0.127	−0.237	−0.333	0.038	0.095	0.989
	coef.	0.082	−0.326	0.033	−0.240	−0.131	−0.246	−0.340	0.025	0.082	0.973
10-10	*z*-value	204.2	−123.5	6.13	−58.2	−88.1	−58.8	−108.9	12.1	14.7	113.8
	.025	0.083	−0.353	0.019	−0.262	−0.179	−0.261	−0.358	0.064	0.077	0.896
	.975	0.085	−0.342	0.037	−0.244	−0.171	−0.244	−0.345	0.089	0.101	0.927
	coef.	0.084	−0.347	0.028	−0.253	−0.175	−0.253	−0.351	0.076	0.089	0.911
10-17	*z*-value	204.7	−110.2	−7.18	−39.1	−82.6	−77.6	−127.1	13.7	10.5	129.7
	.025	0.080	−0.287	−0.043	−0.181	−0.156	−0.327	−0.399	0.069	0.048	0.955
	.975	0.081	−0.277	−0.024	−0.164	−0.149	−0.311	−0.387	0.092	0.071	0.984
	coef.	0.081	−0.282	−0.034	−0.173	−0.152	−0.319	−0.393	0.080	0.059	0.969
10-24	*z*-value	249.9	−130.0	−17.4	−43.5	−89.0	−86.5	−143.8	12.7	−8.89	159.9
	.025	0.084	−0.306	−0.083	−0.187	−0.150	−0.331	−0.411	0.057	−0.056	1.06
	.975	0.085	−0.297	−0.067	−0.171	−0.143	−0.316	−0.400	0.077	−0.036	1.09
	coef.	0.085	−0.301	−0.075	−0.179	−0.146	−0.324	−0.405	0.067	−0.046	1.08
10-31	*z*-value	235.1	−138.7	−38.2	−35.1	−96.4	−93.7	−144.3	10.8	−27.3	189.5
	.025	0.077	−0.298	−0.170	−0.149	−0.149	−0.334	−0.376	0.043	−0.140	1.17
	.975	0.078	−0.290	−0.153	−0.134	−0.143	−0.320	−0.366	0.062	−0.121	1.19
	coef.	0.078	−0.294	−0.162	−0.141	−0.146	−0.327	−0.371	0.052	−0.130	1.18
11-07	*z*-value	291.1	−169.7	−60.8	2.00	−83.9	−119.6	−151.6	25.2	15.2	176.5
	.025	0.080	−0.306	−0.236	0.000	−0.108	−0.351	−0.325	0.096	0.053	0.918
	.975	0.081	−0.299	−0.221	0.014	−0.103	−0.340	−0.316	0.112	0.068	0.939
	coef.	0.081	−0.302	−0.228	0.007	−0.105	−0.345	−0.321	0.104	0.061	0.929
11-14	*z*-value	343.9	−193.5	−72.2	−44.0	−177.7	−114.2	−166.4	29.3	−48.1	256.5
	.025	0.083	−0.307	−0.241	−0.142	−0.212	−0.304	−0.313	0.100	−0.182	1.21
	.975	0.084	−0.301	−0.228	−0.130	−0.207	−0.294	−0.306	0.114	−0.168	1.22
	coef.	0.084	−0.304	−0.235	−0.136	−0.210	−0.299	−0.310	0.107	−0.175	1.21
11-21	*z*-value	394.2	−153.9	−64.5	−39.9	−141.9	−103.2	−159.4	41.6	−22.4	251.5
	.025	0.084	−0.221	−0.195	−0.116	−0.152	−0.250	−0.273	0.136	−0.083	1.10
	.975	0.085	−0.215	−0.183	−0.105	−0.147	−0.241	−0.266	0.149	−0.069	1.12
	coef.	0.084	−0.218	−0.189	−0.111	−0.149	−0.245	−0.269	0.143	−0.076	1.11
11-28	*z*-value	397.2	−105.6	−18.1	−95.7	−134.7	−83.1	−145.9	47.9	2.74	229.8
	.025	0.083	−0.152	−0.054	−0.255	−0.148	−0.206	−0.249	0.166	0.003	1.05
	.975	0.084	−0.147	−0.044	−0.245	−0.144	−0.196	−0.242	0.180	0.017	1.07
	coef.	0.083	−0.150	−0.049	−0.250	−0.146	−0.201	−0.246	0.173	0.010	1.06
12-05	*z*-value	495.5	−96.4	−32.1	−54.4	−96.1	−89.7	−126.8	41.3	47.6	215.6
	.025	0.090	−0.130	−0.089	−0.140	−0.097	−0.205	−0.198	0.136	0.152	0.924
	.975	0.091	−0.125	−0.079	−0.130	−0.093	−0.196	−0.192	0.150	0.166	0.941
	coef.	0.090	−0.127	−0.084	−0.135	−0.095	−0.201	−0.195	0.143	0.159	0.932
12-12	*z*-value	591.9	−56.7	25.4	−104.0	−81.0	−74.2	−117.0	54.6	108.1	177.8
	.025	0.095	−0.073	0.053	−0.231	−0.078	−0.161	−0.170	0.182	0.348	0.747
	.975	0.096	−0.068	0.062	−0.222	−0.075	−0.152	−0.164	0.196	0.361	0.764
	coef.	0.095	−0.071	0.057	−0.226	−0.077	−0.156	−0.167	0.189	0.355	0.756
12-19	*z*-value	649.1	−31.2	64.2	−122.4	10.8	−80.3	−115.8	35.9	159.0	151.3
	.025	0.096	−0.041	0.134	−0.262	0.008	−0.172	−0.166	0.122	0.527	0.647
	.975	0.096	−0.037	0.143	−0.254	0.012	−0.164	−0.161	0.136	0.540	0.664
	coef.	0.096	−0.039	0.139	−0.258	0.010	−0.168	−0.164	0.129	0.533	0.655

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.

the seasonal flu epidemic spanning nearly a decade (nine flu seasons between 2003-2012, See Fig. 2a). These are obtained by looking for Influenza related diagnostic codes reported in each week in each county in the Truven Marketscan insurance claims database (51). This database consists of over 150 million patients *i.e.* almost a third of the US population, and despite limitations (under-reporting of non-severe influenza cases, and reporting/coding uncertainties), provide a detailed record of flu season incidence dynamics. 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 Materials and 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 (52) (See Fig. 2b). We note here that the number of clusters (four) is chosen via standard heuristic considerations (53), 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. 2c). 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 (50) 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. 2d). 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 county with count sequence x as:

$$\widehat{\nu(x)} \triangleq L(x, G^\star) - \widehat{H(X)} \to \mathcal{D}(X||G^\star) \qquad [2]$$

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 a consistent estimate (54) of the entropy rate of the process X from x. This is non-trivial (55, 56) if X is not an iid process. We may either: 1) estimate the entropy rate from the observed sample path (57), 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. For simiplicity we present results with only the second approach (See Fig. 2e), *i.e.* using a lower bound to the UnIT risk, which is nevertheless demonstrated to have significant predictive value, especially when scaled up with the percent of the county-specific urban population.

The estimated urban-UnIT risk obtained by scaling the UnIT risk with the fraction of urban population is then used to verify its dominant explanatory role amomgst suspected covariates as discussed before. Finally this risk phenotype is used to make weekly case count forecasts, one week ahead of time, on a per county basis. The forecast model (See Matherials and Methods) is simple; essentially an ensemble regressor with the urban-UnIT risk as an input feature, along with the previous week's county-wise count totals, which as shown in Fig. 3 still outperforms more complex state of the art approaches. This result suggests that the urban-UnIT risk is more "information-dense" about the underlying stochastic phenomena driving the incidence dynamics.

Limitations & Conclusion. A source of uncertainty in our approach is the use of diagnostic codes from insurance claims to infer seasonal flu incidence. Influenza is in general hard to track, since less severe cases are seldom reported. Additionally, electronic health records are also inherently noisy, and suffers from potential coding errors by physicians, and other artifacts. Similarly the number of confirmed COVID-19 cases is also a function of how many tests are actually administered, and the fraction of the infected population who are asymptomatic. Thus, we are forecasting the number of detected cases as opposed to true disease incidence.

Importantly our results do not imply that Influenza and COVID-19 are similar in their clinical progression. Indeed, a limitation of our approach is its reduced ability to predict COVID-19-related deaths (See SI Fig. 2a). Our death count forecasts are worse than the top few contributors (58) to the COVID-19 forecasthub. We hypothesize that this reduced effectiveness is attributable to differences in the clinical progression of Influenza and COVID-19: COVID-19 is a more serious disease, 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.

In this study we demonstrate that leveraging the knowledge of the incidence fluctuations in one epidemic informs another with a broadly similar transmission mechanism, despite differences in the epidemiological parameters and the disease processes. The COVID-19 pandemic has highlighted the need for tools to forecast case counts early in the course of future pandemics, when only sparse data is available to train upon, by leveraging incidence pertaining to different epidemics of the past.

Data Sharing. Data sources are enumerated in Materials and Methods. With the exception of Truven Marketscan, the sources are in the public domain. Generated models are publicly available at https://github.com/zeroknowledgediscovery/unitcov, which includes the complete forecast software.

Materials and Methods

We begin by describing the forecast model, followed by the mathematical details underlying the risk measure itself.

Forecast Model. The UnIT score (ν) is a spatially varying time-invariant measure. Thus, to forecast 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 forecast made for t at time $t-1$):

$$\text{UnIT risk correction} \qquad X^\star_t = g_t(X_t, \nu, v_1, \cdots, v_m) \qquad [3a]$$

$$\text{Regressor training} \qquad X_t = h_t(X^\star_{t-1}) \qquad [3b]$$

$$\text{Forecasting estimate} \qquad \widehat{X}_{t+1} = h_t(X^\star_t) \qquad [3c]$$

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 Eq. (11)). ν is the urban-UnIT risk, and the rest of the variables $v_1, \cdots, v_m$ (as described in Table 1) 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 ensemble regressor with the UnIT-corrected case count as one of the features/inputs. 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. 3d). 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. 4).

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 (59), the short time series distance (STS)(60), which takes into account of irregularity in sampling rates, the edit based distances(61) with generalizations to continuous sequences(62), and the dynamic time warping (DTW)(63), 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 (49) 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 (64). 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 (65) 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:

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

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

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

Then, by law of large numbers (66), 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 Sample Paths. 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 in the SI text) if we restrict our stochastic processes to those generated by Probabilistic Finite State Automata (PFSA) (67–70). PFSA are semantically succinct and can model discrete-valued stochastic processes of any finite Markov order, and can approximate arbitrary Hidden Markov Models (68) (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:

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

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 (See SI text for pseudocode of algorithm). In our approach, we use the set of four PFSA models shown in SI 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. 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. SI 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 (71).

Probabilistic Finite Automata.

Definition 1 (PFSA). *A probabilistic finite-state automaton G is a quadruple $\left(Q, \Sigma, \delta, \widetilde{\pi}\right)$, 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 (72).

Definition 2 (Observation and Transition Matrices)**.** *Given a PFSA* $\left(Q, \Sigma, \delta, \widetilde{\pi}\right)$*, 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 $\mathbf{p}_G^T \Pi_G = \mathbf{p}_G^T$ (73), 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*

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

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

Definition 4 (Sequence-Induced Distributions)**.** *For a PFSA $G = \left(Q, \Sigma, \delta, \widetilde{\pi}\right)$, 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 $G = \left(Q, \Sigma, \delta, \widetilde{\pi}\right)$ 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 (74). 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*

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

and the KL divergence

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

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 (74) of a PFSA G generating $x \in \Sigma^d$ is given by*

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

Algorithm 1 in the SI text 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*

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

in probability as $d \to \infty$.

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 $a_{ij} = \exp\left(-d_{ij}^2/2\right)$.

Data Source: COVID-19 Incidence & Putative 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 (2020-12-19) 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, based on the 2010 US decennial census.

Data Source: Seasonal Influenza Incidence. The source of incidence counts for seasonal flu epidemic is the Truven MarketScan database (51). 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, and 3) **487.1** Influenza with other respiratory manifestations and 4) **487.8** Influenza with other manifestations

Discretization of 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. Eq. (5). 2) Cluster counties using this similarity measure using standard spectral clustering algorithm (52). 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 (50). 6)

Finally, estimate UnIT risk as

$$\widehat{\nu(x)} \triangleq L(x, G^\star) - \widehat{H(X)} \to \mathcal{D}(X||G^\star) \quad [10]$$

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 (49). 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.

UnIT Correction To Case Count Forecast. We fit a generalized linear model (39, 75) (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:

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k, \quad [11]$$

where $X_1, X_2, \ldots, X_k$ are explanatory variables (covariates). The counts are for all one-week periods between 2020-04-04 to 2020-12-19. 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: 1) *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; and 2) *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 before, 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 in the SI text. 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. 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 for training regressors, we also use the case count vectors and its corrected version produced by Algorithm 2 (See SI text).

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.

1. R Li, et al., Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science* **368**, 489–493 (2020).
2. ND Yanez, NS Weiss, JA Romand, MM Treggiari, Covid-19 mortality risk for older men and women. *BMC Public Heal.* **20**, 1–7 (2020).
3. L Fang, G Karakiulakis, M Roth, Are patients with hypertension and diabetes mellitus at increased risk for covid-19 infection? *The Lancet. Respir. Medicine* **8**, e21 (2020).
4. I Chattopadhyay, E Kiciman, JW Elliott, JL Shaman, A Rzhetsky, Conjunction of factors triggering waves of seasonal influenza. *Elife* **7**, e30756 (2018).
5. MJ Keeling, P Rohani, Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecol. Lett.* **5**, 20–29 (2002).
6. C Viboud, et al., Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447–51 (2006).
7. V Colizza, A Barrat, M Barthelemy, A Vespignani, 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).
8. D Balcan, et al., Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc Natl Acad Sci U S A* **106**, 21484–9 (2009).
9. D Balcan, A Vespignani, Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nat Phys* **7**, 581–586 (2011).
10. RM Eggo, S Cauchemez, NM Ferguson, Spatial dynamics of the 1918 influenza pandemic in england, wales and the united states. *J R Soc Interface* **8**, 233–43 (2011).
11. D Brockmann, D Helbing, The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–42 (2013).
12. J Shaman, M Kohn, Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc Natl Acad Sci U S A* **106**, 3243–8 (2009).
13. G Chowell, et al., The influence of climatic conditions on the transmission dynamics of the 2009 a/h1n1 influenza pandemic in chile. *Bmc Infect. Dis.* **12** (2012).
14. JR Gog, et al., Spatial transmission of 2009 pandemic influenza in the us. *PLoS Comput. Biol* **10**, e1003635 (2014).
15. V Charu, et al., Human mobility and the spatial transmission of influenza in the united states. *PLoS Comput. Biol* **13**, e1005382 (2017).
16. AL Phelan, R Katz, LO Gostin, The novel coronavirus originating in wuhan, china: challenges for global health governance. *Jama* **323**, 709–710 (2020).
17. A Pan, et al., Association of public health interventions with the epidemiology of the covid-19 outbreak in wuhan, china. jama (2020).
18. N Altieri, et al., Curating a covid-19 data repository and forecasting county-level death counts in the united states. *arXiv preprint arXiv:2005.07882* (2020).
19. I COVID, CJ Murray, , 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).
20. D Benvenuto, M Giovanetti, L Vassallo, S Angeletti, M Ciccozzi, Application of the arima model on the covid-2019 epidemic dataset. *Data brief*, 105340 (2020).
21. SJ Fong, G Li, N Dey, RG Crespo, E Herrera-Viedma, Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak. *arXiv preprint arXiv:2003.10776* (2020).
22. G Ding, X Li, Y Shen, J Fan, Brief analysis of the arima model on the covid-19 in italy. *medRxiv* (2020).
23. C for Disease Control, 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).
24. NS Wong, CC Leung, SS Lee, Abrupt subsidence of seasonal influenza after covid-19 outbreak, hong kong, china. *Emerg. Infect. Dis.* **26**, 2752 (2020).
25. SJ Olsen, et al., Decreased influenza activity during the covid-19 pandemic—united states, australia, chile, and south africa, 2020. *Morb. Mortal. Wkly. Rep.* **69**, 1305 (2020).
26. RJJ Soo, CJ Chiew, S Ma, R Pung, V Lee, Decreased influenza incidence under covid-19 control measures, singapore. *Emerg. infectious diseases* **26**, 1933 (2020).
27. M Al-Raeei, 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. *Clin. epidemiology global health* (2020).
28. MA Billah, MM Miah, MN Khan, Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. *PloS one* **15**, e0242128 (2020).
29. S Dharmaratne, et al., Estimation of the basic reproduction number (r0) for the novel coronavirus disease in sri lanka. *Virol. J.* **17**, 1–7 (2020).
30. DP Oran, EJ Topol, Prevalence of asymptomatic sars-cov-2 infection: A narrative review. *Annals Intern. Medicine* (2020).
31. NH Leung, C Xu, DK Ip, BJ Cowling, The fraction of influenza virus infections that are asymptomatic: a systematic review and meta-analysis. *Epidemiol. (Cambridge, Mass.)* **26**, 862 (2015).
32. PW Brady, et al., Trends in covid-19 risk-adjusted mortality rates. *J Hosp Med.* (2020).
33. Y Luo, J Yan, S McClure, Distribution of the environmental and socioeconomic risk factors on covid-19 death rate across continental usa: a spatial nonlinear analysis. *Environ. Sci. Pollut. Res.*, 1–13 (2020).
34. CH Zhang, GG Schwartz, Spatial disparities in coronavirus incidence and mortality in the united states: an ecological analysis as of may 2020. *The J. Rural. Heal.* **36**, 433–445 (2020).
35. R Khazanchi, et al., County-level association of social vulnerability with covid-19 cases and deaths in the usa. *J. general internal medicine* **35**, 2784–2787 (2020).
36. A Ehlert, The socioeconomic determinants of covid-19: A spatial analysis of german county level data. *medRxiv* (2020).
37. A Mollalo, B Vahedi, KM Rivera, Gis-based spatial modeling of covid-19 incidence rate in the continental united states. *Sci. The Total. Environ.* **728**, 138884 (2020).
38. F Sun, SA Matthews, TC Yang, MH Hu, A spatial analysis of covid-19 period prevalence in us counties through june 28, 2020: Where geography matters? *Annals Epidemiol.* (year?).
39. D Hedeker, R Gibbons, *Longitudinal data analysis*, Wiley series in probability and statistics. (Wiley-Interscience, Hoboken, N.J.), (2006).

40. AL Bertozzi, E Franco, G Mohler, MB Short, D Sledge, The challenges of modeling and forecasting the spread of covid-19. *Proc. Natl. Acad. Sci.* **117**, 16732–16738 (2020).
41. A Arenas, et al., A mathematical model for the spatiotemporal epidemic spreading of covid19. *MedRxiv* (2020).
42. L Li, et al., Propagation analysis and prediction of the covid-19. *Infect. Dis. Model.* **5**, 282–292 (2020).
43. Y Contoyiannis, et al., A universal physics-based model describing covid-19 dynamics in europe. *Int. J. Environ. Res. Public Heal.* **17**, 6525 (2020).
44. EL Ray, et al., Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *MedRXiv* (2020).
45. Á Briz-Redón, Á Serrano-Aroca, A spatio-temporal analysis for exploring the effect of temperature on covid-19 early evolution in spain. *Sci. Total. Environ.*, 138811 (2020).
46. SJ Kim, W Bostwick, Social vulnerability and racial inequality in covid-19 deaths in chicago. *Heal. education & behavior* **47** (2020).
47. J Cordes, MC Castro, Spatial analysis of covid-19 clusters and contextual factors in new york city. *Spatial Spatio-temporal Epidemiol.* **34**, 100355 (2020).
48. D Donoho, 50 years of data science. *J. Comput. Graph. Stat.* **26**, 745–766 (2017).
49. TM Cover, JA Thomas, *Elements of Information Theory*. (Wiley-Interscience, New York, NY, USA), (1991).
50. I Chattopadhyay, H Lipson, Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013).
51. L Hansen, The truven health marketscan databases for life sciences researchers. *Truven Heal. Ananlytics IBM Watson Heal.* (2017).
52. AY Ng, MI Jordan, Y Weiss, On spectral clustering: Analysis and an algorithm in *Advances in neural information processing systems*. pp. 849–856 (2002).
53. M Alshammari, M Takatsuka, Approximate spectral clustering with eigenvector selection and self-tuned k. *Pattern Recognit. Lett.* **122**, 31–37 (2019).
54. E Sober, Likelihood and convergence. *Philos. Sci.* **55**, 228–237 (1988).
55. T Schürmann, P Grassberger, Entropy estimation of symbol sequences. *Chaos: An Interdiscip. J. Nonlinear Sci.* **6**, 414–427 (1996).
56. P Grassberger, Estimating the information content of symbol sequences and efficient codes. *IEEE Transactions on Inf. Theory* **35**, 669–675 (1989).
57. I Chattopadhyay, H Lipson, Computing entropy rate of symbol sources & a distribution-free limit theorem in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*. (IEEE), pp. 1–6 (2014).
58. N Reich, Viz - covid-19 forecast hub — covid-19 (https://viz.covid19forecasthub.org/) (2020) (Accessed on 11/29/2020).
59. J Lin, E Keogh, S Lonardi, B Chiu, 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*. (ACM), pp. 2–11 (2003).
60. CS Möller-Levet, F Klawonn, KH Cho, O Wolkenhauer, Fuzzy clustering of short time-series and unevenly distributed sampling points in *International Symposium on Intelligent Data Analysis*. (Springer), pp. 330–340 (2003).
61. G Navarro, A guided tour to approximate string matching. *ACM computing surveys (CSUR)* **33**, 31–88 (2001).
62. L Chen, MT Özsu, V Oria, Robust and fast similarity search for moving object trajectories in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. (ACM), pp. 491–502 (2005).
63. F Petitjean, A Ketterlin, P Gançarski, A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognit.* **44**, 678–693 (2011).
64. CW Helstrom, *Probability and stochastic processes for engineers*. (Macmillan Coll Division), (1991).
65. RW Hamming, Error detecting and error correcting codes. *The Bell system technical journal* **29**, 147–160 (1950).
66. FM Dekking, C Kraaikamp, HP Lopuhaä, LE Meester, *A Modern Introduction to Probability and Statistics: Understanding why and how*. (Springer Science & Business Media), (2005).
67. JP Crutchfield, The calculi of emergence: computation, dynamics and induction. *Phys. D: Nonlinear Phenom.* **75**, 11–54 (1994).
68. P Dupont, F Denis, Y Esposito, Links between probabilistic automata and hidden markov models: probability distributions, learning models and induction algorithms. *Pattern recognition* **38**, 1349–1371 (2005).
69. I Chattopadhyay, H Lipson, Data smashing: uncovering lurking order in data. *J. The Royal Soc. Interface* **11**, 20140826 (2014).
70. I Chattopadhyay, Causality networks. *arXiv preprint arXiv:1406.6651* (2014).
71. HA Dau, et al., The ucr time series classification archive (2018) https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.
72. J Bondy, U Murty, Graph theory (2008). *Grad. Texts Math* (2008).
73. M Vidyasagar, *Hidden markov processes: Theory and applications to biology*. (Princeton University Press) Vol. 44, (2014).
74. TM Cover, JA Thomas, *Elements of information theory*. (John Wiley & Sons), (2012).
75. WH Greene, *Econometric analysis*. (Pearson Education India), (2003).

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

List of Figures

List of Tables

List of Algorithms

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

	Team name	Description	Home
1	Columbia University (Abbr. CU-select)	A metapopulation county-level SEIR model for projecting future COVID-19 incidence and deaths. This forecast is the scenario we believe to be most plausible given the current setting.	https://blogs.cuit.columbia.edu/jls106/publications/covid-19-findings-simulations/
2	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/
3	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
4	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
5	COVID-19 Forecast Hub (Abbr. COVIDhub-baseline)	This model is a baseline predictive model.	https://covid19forecasthub.org/
6	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
7	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/
8	QJHong (Abbr. QJHong-Encounter)	today's "Daily New Confirmed Cases" + today's "Encounter Density" $\Rightarrow$ today's newly infected Cases $\Rightarrow$ next 2-3 weeks' "Daily New Confirmed Cases"	https://qjhong.github.io
9	LockNQuay (Abbr. LNQ-ens1)	County-level ensemble of boosted tree and neural net models. Lots of engineered features.	https://www.kaggle.com/sasrdw/locknquay
10	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/
11	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

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`

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 Π_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* σ *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 $\mathbf{p}_{\text{new}} / \left\|\mathbf{p}_{\text{new}}\right\|_1$;
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_{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.

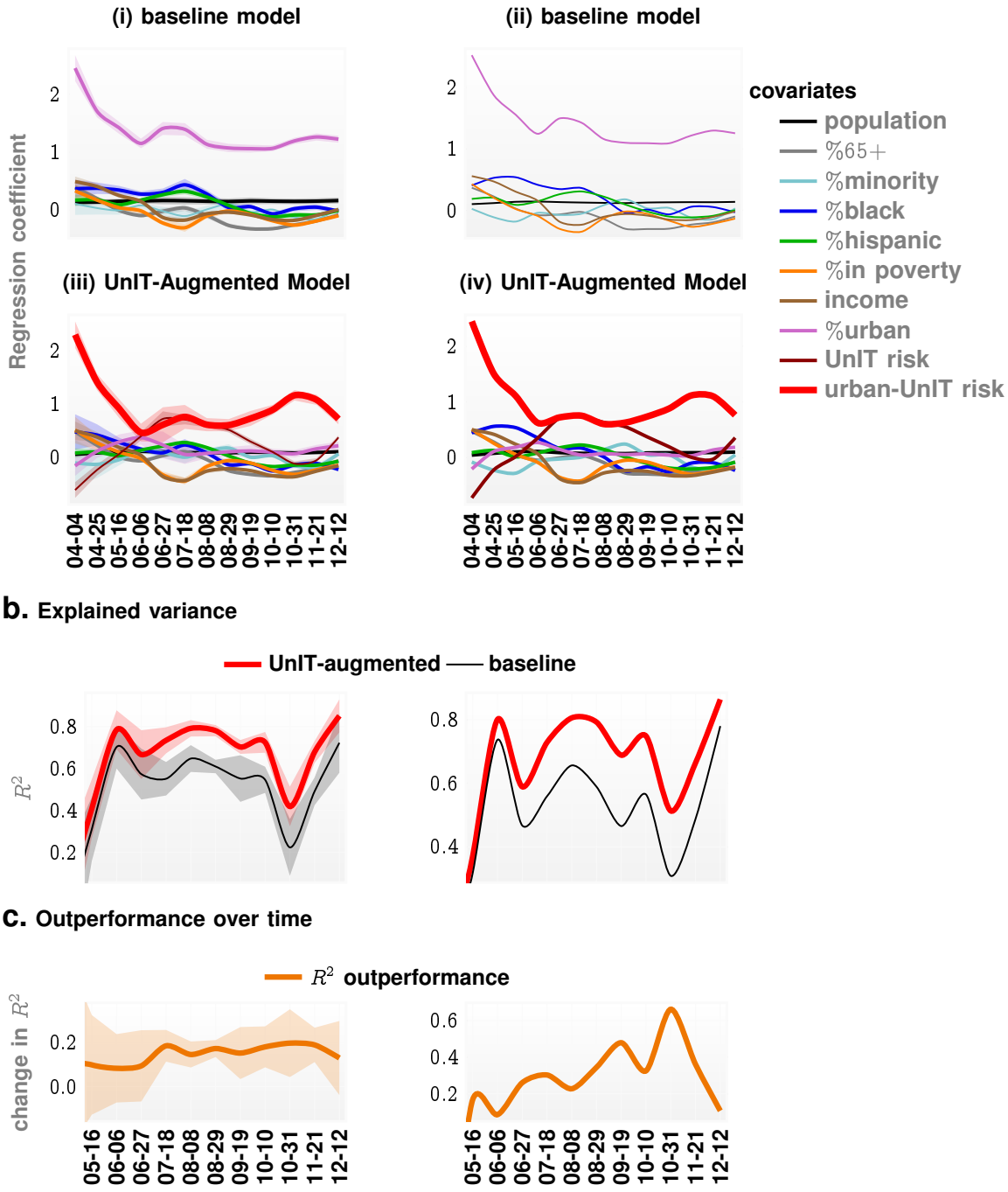
for *each regressor* $\texttt{Regr} \in \mathcal{R}$ **do**
    Let $X_{\text{train}} = [D_{t-1}, C_{t-1}, D_{glm,t-1}, D_{glm,t}, C_{glm,t-1}, C_{glm,t}]$;
    Let $y_{\text{train}} = D_t$;
    Fit `Regr` with $X_{\text{train}}, y_{\text{train}}$;
    Let $X_{\text{pred}} = [D_t, C_t, D_{glm,t-1}, D_{glm,t}, C_{glm,t-1}, C_{glm,t}]$;
    Let $y_{\text{pred},\texttt{Regr}}$ be the prediction `Regr` makes with X_{pred};
Let $D_{\text{pred},t+1} = \sum_{r \in \mathcal{R}} y_{\text{pred},r} / |\mathcal{R}|$;
return $D_{\text{pred},t+1}$;

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

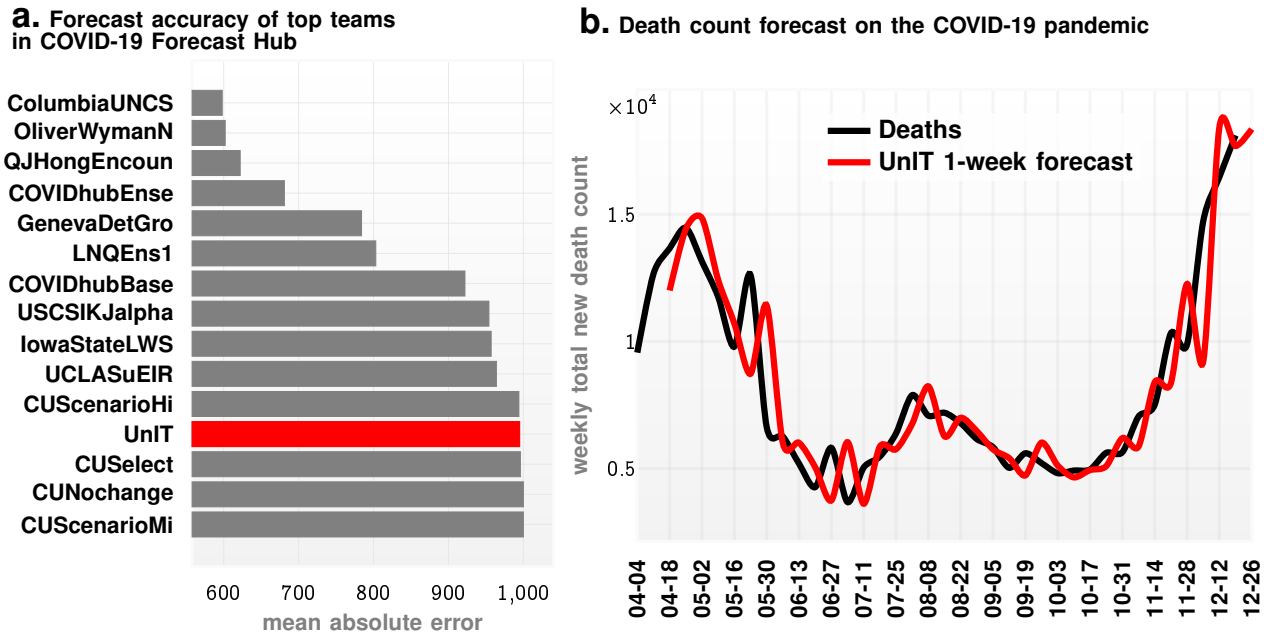
SI Data Tab. II
COEFFICIENTS INFERRED IN MULTI-VARIATE REGRESSION FOR WEEKLY COVID-19-RELATED DEATH TOTALS

		pop	%65+	%minority	%black	%hispanic	%poverty	income	%urban	pre-UnIT	UnIT
09-12	z-value	25.4	9.09	0.969	2.98	16.2	−1.60	−10.7	−2.44	14.2	10.0
	.025	0.068	0.140	−0.044	0.041	0.192	−0.116	−0.347	−0.258	0.638	0.550
	.975	0.079	0.217	0.130	0.202	0.245	0.012	−0.240	−0.028	0.842	0.818
	coef.	0.073	0.178	0.043	0.121	0.219	−0.052	−0.294	−0.143	0.740	0.684
09-19	z-value	22.3	9.84	6.71	−1.40	14.8	−1.91	−10.8	0.018	16.6	7.50
	.025	0.060	0.143	0.167	−0.109	0.166	−0.117	−0.315	−0.105	0.696	0.353
	.975	0.071	0.214	0.305	0.018	0.217	0.002	−0.219	0.107	0.882	0.603
	coef.	0.066	0.178	0.236	−0.045	0.191	−0.058	−0.267	0.001	0.789	0.478
09-26	z-value	26.4	10.9	3.48	−1.13	11.0	−2.91	−13.2	1.14	11.9	9.94
	.025	0.071	0.162	0.060	−0.114	0.122	−0.159	−0.409	−0.044	0.489	0.521
	.975	0.082	0.233	0.215	0.030	0.175	−0.031	−0.303	0.168	0.683	0.776
	coef.	0.076	0.197	0.138	−0.042	0.148	−0.095	−0.356	0.062	0.586	0.648
10-03	z-value	23.2	5.13	2.84	−2.41	7.22	−0.941	−9.81	0.512	10.1	10.1
	.025	0.067	0.063	0.034	−0.159	0.074	−0.096	−0.314	−0.078	0.402	0.532
	.975	0.079	0.140	0.187	−0.017	0.130	0.034	−0.209	0.134	0.595	0.790
	coef.	0.073	0.101	0.111	−0.088	0.102	−0.031	−0.262	0.028	0.498	0.661
10-10	z-value	22.3	2.76	−0.075	−1.17	4.64	−0.398	−9.59	1.42	9.97	9.72
	.025	0.067	0.016	−0.084	−0.121	0.038	−0.077	−0.308	−0.029	0.387	0.503
	.975	0.080	0.094	0.078	0.030	0.092	0.051	−0.204	0.179	0.577	0.757
	coef.	0.073	0.055	−0.003	−0.045	0.065	−0.013	−0.256	0.075	0.482	0.630
10-17	z-value	20.9	8.84	1.32	−1.51	5.30	−6.57	−13.6	−1.39	7.72	13.9
	.025	0.064	0.124	−0.027	−0.135	0.048	−0.296	−0.433	−0.164	0.257	0.720
	.975	0.077	0.194	0.137	0.018	0.104	−0.160	−0.324	0.028	0.433	0.957
	coef.	0.071	0.159	0.055	−0.059	0.076	−0.228	−0.378	−0.068	0.345	0.839
10-24	z-value	25.4	3.58	4.46	−5.79	−2.99	−8.06	−15.1	5.31	9.92	8.26
	.025	0.077	0.029	0.087	−0.254	−0.072	−0.330	−0.445	0.155	0.338	0.367
	.975	0.090	0.098	0.223	−0.126	−0.015	−0.201	−0.343	0.337	0.505	0.596
	coef.	0.084	0.063	0.155	−0.190	−0.044	−0.266	−0.394	0.246	0.422	0.482
10-31	z-value	22.3	−3.03	2.93	−8.27	−6.91	−7.10	−15.7	2.51	8.76	11.0
	.025	0.070	−0.094	0.033	−0.337	−0.130	−0.296	−0.461	0.025	0.283	0.518
	.975	0.084	−0.020	0.169	−0.208	−0.072	−0.168	−0.358	0.205	0.446	0.743
	coef.	0.077	−0.057	0.101	−0.272	−0.101	−0.232	−0.409	0.115	0.365	0.631
11-07	z-value	32.9	−5.79	−2.52	−1.06	−8.44	−3.77	−13.2	0.622	7.83	13.7
	.025	0.088	−0.137	−0.159	−0.099	−0.137	−0.162	−0.349	−0.055	0.217	0.595
	.975	0.099	−0.068	−0.020	0.029	−0.085	−0.051	−0.259	0.105	0.362	0.794
	coef.	0.093	−0.102	−0.089	−0.035	−0.111	−0.106	−0.304	0.025	0.289	0.695
11-14	z-value	27.7	−5.18	−3.60	−1.70	−7.67	−8.78	−16.7	1.43	6.18	15.9
	.025	0.076	−0.119	−0.205	−0.126	−0.122	−0.309	−0.428	−0.021	0.153	0.692
	.975	0.088	−0.053	−0.060	0.009	−0.072	−0.196	−0.338	0.133	0.295	0.886
	coef.	0.082	−0.086	−0.133	−0.058	−0.097	−0.253	−0.383	0.056	0.224	0.789
11-21	z-value	32.3	−4.54	−5.64	−0.994	−10.1	−9.12	−18.6	5.05	4.24	18.8
	.025	0.078	−0.090	−0.247	−0.089	−0.130	−0.272	−0.401	0.103	0.072	0.721
	.975	0.088	−0.036	−0.120	0.029	−0.088	−0.176	−0.324	0.233	0.196	0.889
	coef.	0.083	−0.063	−0.183	−0.030	−0.109	−0.224	−0.363	0.168	0.134	0.805
11-28	z-value	39.2	−1.46	−4.23	−3.73	−10.8	−10.1	−18.7	3.98	3.95	19.6
	.025	0.089	−0.047	−0.200	−0.173	−0.145	−0.309	−0.412	0.068	0.064	0.764
	.975	0.098	0.007	−0.073	−0.054	−0.100	−0.208	−0.334	0.200	0.189	0.934
	coef.	0.094	−0.020	−0.137	−0.114	−0.123	−0.259	−0.373	0.134	0.127	0.849
12-05	z-value	40.8	−0.910	−4.62	−6.31	−14.5	−9.71	−20.6	7.74	7.12	21.2
	.025	0.081	−0.032	−0.170	−0.202	−0.152	−0.243	−0.358	0.166	0.140	0.706
	.975	0.089	0.012	−0.069	−0.106	−0.116	−0.161	−0.296	0.278	0.247	0.851
	coef.	0.085	−0.010	−0.120	−0.154	−0.134	−0.202	−0.327	0.222	0.193	0.778
12-12	z-value	51.4	2.19	−6.38	−5.04	−15.0	−10.1	−18.9	9.60	6.83	21.6
	.025	0.091	0.002	−0.210	−0.167	−0.152	−0.239	−0.309	0.206	0.125	0.681
	.975	0.099	0.043	−0.111	−0.073	−0.117	−0.161	−0.251	0.311	0.226	0.817
	coef.	0.095	0.023	−0.161	−0.120	−0.134	−0.200	−0.280	0.258	0.176	0.749
12-19	z-value	59.2	5.56	−4.91	−7.95	−12.1	−6.10	−16.9	9.37	9.60	21.8
	.025	0.093	0.036	−0.158	−0.217	−0.116	−0.150	−0.260	0.198	0.194	0.674
	.975	0.099	0.074	−0.068	−0.131	−0.084	−0.077	−0.206	0.302	0.294	0.808
	coef.	0.096	0.055	−0.113	−0.174	−0.100	−0.114	−0.233	0.250	0.244	0.741

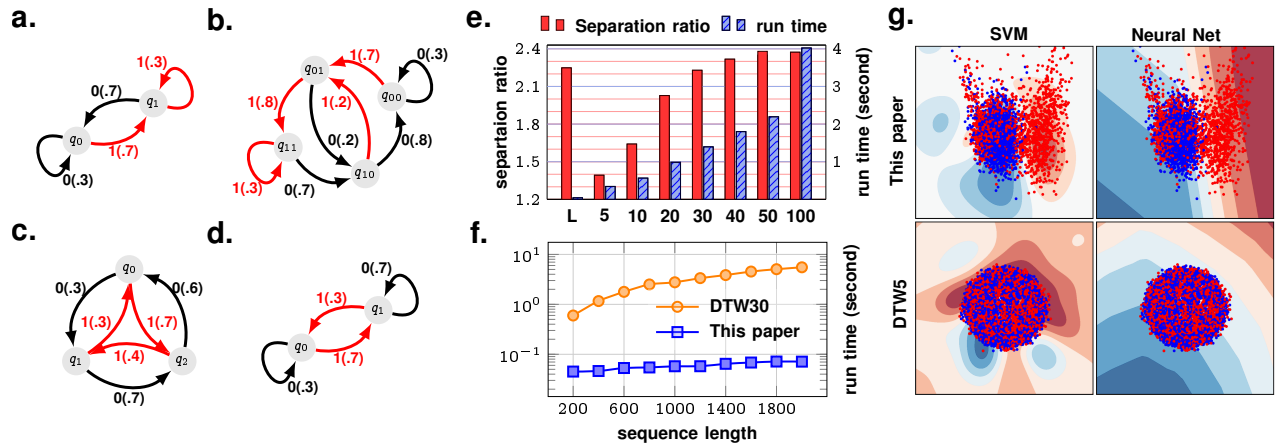
Coefficients with p-value in $[0.01, 0.05)$ are colored blue, and those with p-value ≥ 0.05, red. All other p-values are < 0.01.



SI Data Fig. 1. To test the robustness of the UnIT score as a key influencing variable, we tested two perturbation modes: (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 1 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.



SI 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.



SI Data Fig. 3. **Panel a-d.** Four pre-specified PFSAs to estimate similarity between stochastic sample paths (See (5) in main text). An edge connecting state q to q' is labeled as $\sigma\,(\widetilde{\pi}(q,\sigma))$ 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 the length of the input sequence 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 (1) with decision boundaries obtained by using Support Vector Machines (SVM) and neural networks respectively trained with features constructed from the corresponding dissimilarity measures. The SLD approach yields significantly improved separation.

SI Data Tab. III
COEFFICIENTS IN MULTI-VARIATE REGRESSION FOR COVID-19-RELATED DEATH COUNT TOTAL AS OF 2020-12-19

	coef.	z-value	.025	.975
pop	0.072	193.237	0.071	0.073
%65+	0.216	91.050	0.211	0.220
%minority	0.090	18.686	0.081	0.100
%black	0.038	8.505	0.029	0.047
%hispanic	0.057	30.005	0.053	0.061
%poverty	0.109	26.494	0.101	0.117
income	0.005	1.789	−0.000	0.011
%urban	−0.005	−0.679	−0.019	0.009
UnIT	0.191	27.011	0.177	0.204
urban UnIT	1.073	120.476	1.055	1.090

All p-values are < 0.0005.

Theorem 1 (Proof of 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

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

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 (2), we have

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

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

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

Because the process generated by G is ergodic, we have

$$\lim_{n \to \infty} A_{x,n} = \lim_{d \to \infty} D_d = \mathcal{D}\left(G \,\|\, G'\right). \tag{1}$$

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

REFERENCES

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

[2] G. Hardy, *Éditions Jacques Gabay, Sceaux* (1992).