When Data Disappear: Public Health Pays As Policy Strays

Thomas McAndrew*,[1] Andrew A. Lover,[2] Garrik Hoyt,[3] and Maimuna S. Majumder[4,5]

[1] *Department of Biostatistics and Health Data Science, College of Health, Lehigh University, Bethlehem, Pennsylvania, United States of America**

[2] *Dept. of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts Amherst, Amherst, Massachusetts, United States of America*

[3] *Dept. of Computer Science and Engineering, PC Rossin College of Engineering and Applied Sciences, Lehigh University, Bethlehem, Pennsylvania, United States of America*

[4] *Harvard Medical School, Boston, MA, United States of America*

[5] *Boston Children's Hospital, Boston, MA, United States of America*

(Dated: February 4, 2025)

On January 20th, 2025, President Trump signed multiple executive orders that greatly limit the ability of public health entities like the Centers for Disease Control and Prevention (CDC)—and Health and Human Services (HHS) more broadly—to release critical public health data. [1] Executive orders (EOs) are not new, and have been used across many presidential administrations to shape policy at federal-, state-, and local-levels.

However, last month's EOs targeted public health infrastructure, leading to unspecified delays in the release of many routine CDC data sources and reports, including the *Morbidity and Mortality Weekly Report*, which was—until the week of January 23, 2025—in continual circulation since 1952. Public health decision-making relies on real-time data and statistical models to quickly identify and quantify risks present in populations, to inform the public in a timely fashion, and to deliver targeted health programming where required. [2–4] However, there are currently no specific statutes that mandate the collection or reporting of many vital epidemiological data sources.

Public health data builds policy

Infectious disease models at national- and state-levels rely almost exclusively on government-maintained surveillance data.[5] When the influenza season begins, accurate and timely forecasts complement Advisory Committee on Immunization Practices (ACIP) recommendations by providing public health officials real-time and forward-looking information about the season: when will influenza transmission peak? How many cases will be reported? How many will be hospitalized? Forecasts can also help assess the impact of preventative actions: when is the right time to hold vaccine clinics? in what locations? when should we alert hospitals that the season has started? Modeling—backed by the several governmentally-supported epidemiological data sources—give public health officials access to a useful, albeit uncertain, future.[6]

To illustrate the importance of these vital epidemiological data sources that directly support public health decision-making, let us rewind time to October 2023 (the beginning of the official 2023/24 influenza reporting season), and directly compare influenza forecasts of US incident hospitalizations using two transmission models: one model that takes advantage of several epidemiological data sources routinely collected and disseminated (until these recent EOs) and a second model that uses only one of the several data sources. Influenza is a benchmark pathogen with a robust collection of data, making it an ideal use case for illustrating the importance of public health data.

The Blue model uses a diverse suite of epidemiological data sources, including the NHSN dataset (containing (1) weekly incident influenza hospitalizations and (2) the percentage of hospitals that have reported data); ILI-NET (that hosts (3) weekly percent influenza-like illness [ILI] across providers); MMWR (which reports,

* mcandrew@lehigh.edu

among other data, (4) vaccine effectiveness); and the GHCd dataset (hosted by NOAA, which reports (5) average temperature and (6) barometric pressure—both of which modulate influenza transmission). [7–9] We also collected (7) the estimated population size from the most recent US census which means the Blue model relies on seven data sets. In contrast, The Red model solely relies on the NHSN hospitalization data. CMS (Centers for Medicare & Medicaid Services) required that this NHSN dataset be collected, potentially making it less prone to pause. In addition, all statistical models of influenza hospitalization train on this data source. All of the aforementioned data sources are federally supported and have been impacted in some way by recent EOs.

To measure the impact of missing data sources, we compare seasonal trajectories by producing parallel 32-week (full season) forecasts of US national-level incident influenza hospitalizations (See Figure 1). Technical details are in the supplement.

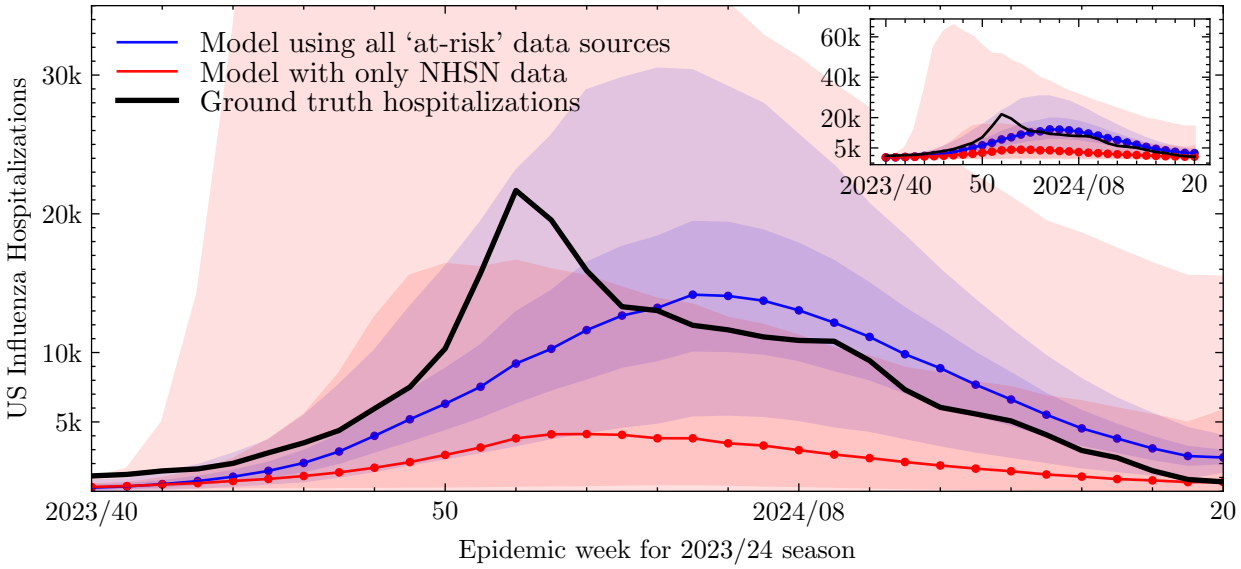

FIG. (1) A 32-week-ahead (full season) forecast of US incident influenza hospitalizations for a model trained on all data sources that are at-risk for pause or removal (blue), and a second model trained only on NHSN hospitalization data (red). Forecasts are represented as a median (solid line and circles), 50% prediction interval (darker shaded area) and 95% prediction interval (lighter shaded area). Ground truth hospitalization data for the (unseen to the models) 2023/24 season in black.

Given the red forecast, using just a single—albeit meaningful—data source, public health officials would have been faced with serious uncertainty about the trajectory of the influenza season, far exceeding the bounds of the plot (see Figure inset). Worse, they may have interpreted the median prediction (solid red line) without considering the associated uncertainty and might have anticipated a mild season, when in fact, it was one of the most intense seasons recorded since the COVID-19 pandemic. This forecast would have likely translated into policy actions that may not have sufficiently emphasized the need for behavioral change communication (BCC) and vaccination, and may have even resulted in mis-communication to hospitals regarding their staffing needs, provision of ventilators and antivirals, and policies to protect staff. A comparison of the Blue model versus Red model forecasts (Figure 1) clearly illustrates the importance of complementary data sources in producing a forecast with sufficient precision for public health agencies to implement proactive health programs.

If the federal government were to no longer collect or maintain the public health datasets included in the Blue Model then we may consequently see a drastic increase in influenza burden. A typical influenza season leads to on average 400k hospitalizations, 20k deaths, all at a cost of approximately $11 billion. Without data from NHSN, ILI-NET, NOAA, and MMWR, modeling efforts to directly support resource allocation,

public health policy, and decision making will be gravely hindered. Moreover, public health data collection is a matter of national security, as highlighted post 9-11. [10]

Public health data prepares us for pandemics

Notably, the current ongoing 'pause' of data collection, analysis, and dissemination is occurring in the midst of a unprecedented national outbreak of H5N1 influenza ("bird flu")—a pathogen with considerable pandemic potential—in humans, birds, and cattle. [11] The fettered ability of the CDC to collect and distribute data, and the inability to allow the CDC to inform the public about this ongoing outbreak, could have dire consequences, nationally and globally.

In stark contrast to the current public health landscape, during the 2009 H1N1 "swine flu" pandemic, the CDC rapidly established an Emergency Operations Center (EOC), which collected, analyzed and rapidly disseminated data and subsequent guidance about the crisis as it unfolded.[12]

Many of the actions taken during the 2009 H1N1 pandemic, and early SARS-CoV-2 responses were coordinated under the National Strategy For Pandemic Influenza Implementation Plan, developed under President Bush's administration in 2006.[13] At its core, this response plan relies on data collection, analysis, and communication—all of which have been hindered by last month's executive orders.

Potential paths forward

Safeguarding important health data sources, especially in the face of ongoing executive orders, requires proactivity. While a wide range of actions are needed, we consider here a non-exhaustive list of realistic and feasible approaches.

Public health data as public good. We must ensure that public health data is prioritized as a public good. To be clear, public health data meets the requirements of a public good: use of health data does not exclude, or reduce availability of this data to another. However, through executive action, public health data can be limited or otherwise suppressed. Congressional processes that designate public health data as a public good could stymie executive actions to remove these data sources.

Public health data at sub-national levels. Academia, industry, local government, and health partners must expand efforts to ensure local control of governmentally-hosted datasets. For influenza forecasting, this would entail, at minimum, storing and managing the collection of the above datasets: NHSN, ILI-NET, and data collated in MMWR reports. Storage of these datasets is a marginal issue. A larger burden than storage is supporting and coordinating collection. Thankfully, resources already exist to this end [14] which include efforts to harmonize data formats to minimize unnecessary manipulations. [15]

Proxy data for forecasting. In preparation for future incidences where public health data disappears, there must be a coordinated effort to collect data sets that can serve as proxies—that is, readily-available data that can approximate unavailable data sources. For example, for influenza, past work demonstrated the importance of human mobility, social media trends, genetic epidemiology, and OTC medicine sales to forecast influenza. However, there is currently no single, organized approach to collect these proxy datasets, nor an accepted method to measure their ability to substitute for traditional influenza signals, such as weekly ILI and incident hospitalizations. By far, the most concerted effort is the Delphi Epidata API. This API has built a unified framework for collecting epidemiological data, already collecting real-time data on COVID-19, dengue, norovirus, and influenza.

Concluding thoughts

Open, readily-available public health data provide reproducible and transparent analyses and interpretations to promote the health and well-being of the US. They serve as a common ground to discuss how to change policy to better serve our citizens. Public health data is a public good, and thus, we must strive towards a future in which it is never subject to removal or interruption—a future in which it is protected, if needed, from executive action.

[1] United States Office of Personnel Management. Initial guidance regarding President Trump's Executive Order defending women, 2025. Accessed: 2025-01-31.

[2] Faruque Ahmed, Jonathan L Temte, Doug Campos-Outcalt, Holger J Schünemann, ACIP Evidence Based Recommendations Work Group, EBRWG, et al. Methods for developing evidence-based recommendations by the Advisory Committee on Immunization Practices (ACIP) of the US Centers for Disease Control and Prevention (CDC). *Vaccine*, 29(49):9171–9176, 2011.

[3] Leah S Fischer. CDC grand rounds: modeling and public health decision-making. *MMWR. Morbidity and Mortality Weekly Report*, 65, 2016.

[4] Chelsea S Lutz, Mimi P Huynh, Monica Schroeder, Sophia Anyatonwu, F Scott Dahlgren, Gregory Danyluk, Danielle Fernandez, Sharon K Greene, Nodar Kipshidze, Leann Liu, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, 19:1–12, 2019.

[5] Sarabeth M Mathis, Alexander E Webber, Tomás M León, Erin L Murray, Monica Sun, Lauren A White, Logan C Brooks, Alden Green, Addison J Hu, Roni Rosenfeld, et al. Evaluation of flusight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations. *Nature communications*, 15(1):6289, 2024.

[6] Sara L Loo, Emily Howerton, Lucie Contamin, Claire P Smith, Rebecca K Borchering, Luke C Mullany, Samantha Bents, Erica Carcelen, Sung-mok Jung, Tiffany Bogich, et al. The US COVID-19 and influenza scenario modeling hubs: delivering long-term projections to guide policy. *Epidemics*, 46:100738, 2024.

[7] Health and Human Services. Weekly hospital respiratory data, 2025.

[8] Centers for Disease Control and Prevention. Outpatient respiratory illness activity map determined by data reported to ILINet, 2025.

[9] National Centers for Environmental Information. Global historical climatology network daily (ghcnd), 2025.

[10] Ali S Khan. Public health preparedness and response in the USA since 9/11: a national health security imperative. *The Lancet*, 378(9794):953–956, 2011.

[11] Thomas P Peacock, Louise Moncla, Gytis Dudas, David VanInsberghe, Ksenia Sukhova, James O Lloyd-Smith, Michael Worobey, Anice C Lowen, and Martha I Nelson. The global H5N1 influenza panzootic in mammals. *Nature*, 637(8045):304–313, 2025.

[12] Michael A Jhung, David Swerdlow, Sonja J Olsen, Daniel Jernigan, Matthew Biggerstaff, Laurie Kamimoto, Krista Kniss, Carrie Reed, Alicia Fry, Lynnette Brammer, et al. Epidemiology of 2009 pandemic influenza A (H1N1) in the United States. *Clinical Infectious Diseases*, 52(suppl_1):S13–S26, 2011.

[13] Nancy J Knauer. The federal response to COVID-19: Lessons from the pandemic. *Hastings LJ*, 73:49, 2022.

[14] Association of State and Territorial Health Officials. Association of State and Territorial Health Officials.

[15] Centers for Disease Control and Prevention. Data modernization initiative.

Technical Document—When Data Disappear: Public Health Pays As Policy Strays

I. DATASETS COLLECTED

To calibrate our model, we collected seven data sets: (1) the COVID-19 Reported Patient Impact and Hospital Capacity (which contains information about influenza) from the National Hospital Safety Network (NHSN). From this dataset we collected hospitalization data as well as (2) data on the percentage of hospitals or facilities reporting data. (3) Public health lab data and (4) clinical lab data from the Outpatient Illness and Viral Surveillance dataset called 'ILI-NET' which was accessed via FluView [1, 2]. (5) The Morbidity and Mortality Weekly Reports which contains interim estimates of vaccine effectiveness against influenza. [3]. These data sources are hosted by the Centers for Disease Control and Prevention (CDC). We also collected (6) the Global Historical Climatology Network daily (GHCd) dataset hosted by the National Oceanic and Atmospheric Administration (NOAA). [4]. In addition, we used the census (7) to estimate the number of individuals living in the United States.

The NHSN dataset was collected from 2021 to 2024 (three influenza seasons). From this dataset we collected state-level data about the weekly number of hospitalizations, including US national hospitalization computed as the sum over states, and the percentage of hospitals or facilities reporting to this dataset. Because in the beginning of the influenza season not all facilities may report data, we estimated the number of hospitalizations as the reported number divided by the percent of facilities reporting.

Influenza-like illness (ILI) data was collected from 2015 to 2024 (seven seasons). [5–8] From the ILI-NET public health lab dataset, we collected state-level, weekly, number of patients who were diagnosed with influenza-like illness and number of patients who attended a healthcare facility for any reason. ILI is a syndromic diagnosis and defined approximately as a patient who is admitted to a healthcare facility with a fever (above 38C) plus cough or sore throat. The percent ILI is defined as those diagnosed with ILI divided by total number of patients. It should be noted that ILI includes a numerous number of respiratory illnesses other than influenza. This is the reason for collecting the clinical lab dataset.

From the ILI-NET clinical labs dataset, we collected state-level, weekly, lab-confirmed percent positive cases of influenza from the same time span: 2015 to 2024. Given one state and one week during the influenza season, we multiply the percent positive for influenza from this lab dataset by number of ILI reports to compute ILI+. [9] ILI+ is an estimate of the number of patients who have confirmed influenza.

We collected GHCd data from 2015 to 2024 to overlap with ILI data. From GHCd, we collected the average weekly temperature and average weekly pressure from the three largest cities in each state (See Figure 2). We defined state level temperature as the average over these three cities and the US national average as the average over the three largest cities in all states. [10]

For our treatment of the model and the data below, we will refer to the set of seasons with a capital S and one season using s and the set of all epidemic weeks as T and one week as t. Though the susceptible disease state is also defined with a capital S there should not be confusion in the treatment of the model below.

I.1. Data pre-processing

A Gaussian filter with standard deviation 2 was applied to the mean (over cities) ambient temperature and barometric pressure data. [11] As the transmission rate will depend on GHCd, and because we do not expect rapid changes in contact patterns, the filter was used to produce a time-dependent transmission rate that changed relatively slowly over time (See Figure 5 and associated section on the reproduction number). Though this was not done, cross-validation could be used to find an optimal standard deviation, comparing a forecast evaluation metric for different standard deviation values. The filter was implemented in Scipy. [12]

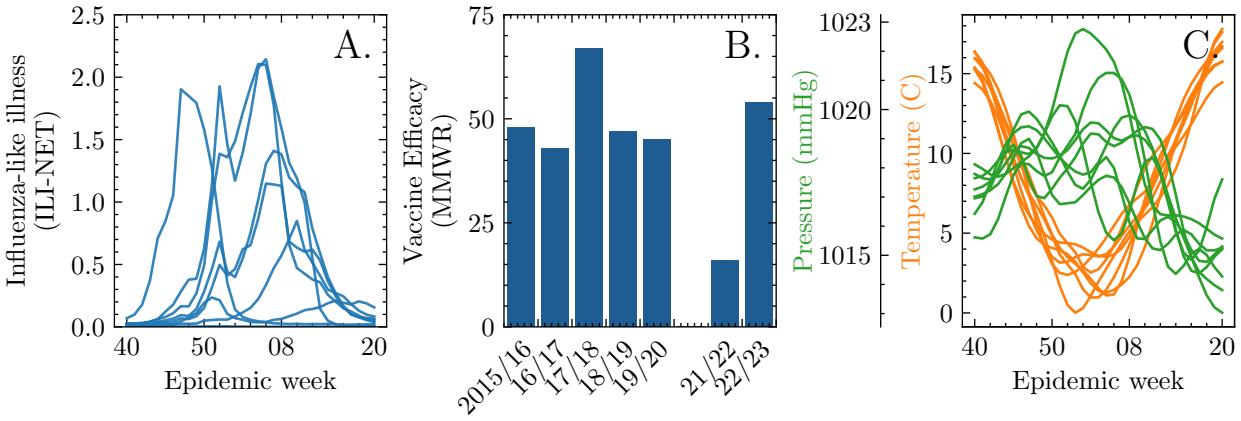

FIG. (1) Data used for modeling collected for the 2015/16 - 2022/23 seasons. (A.) Percent influenza-like illness that resulted in lab-confirmed influenza (i.e. ILI+) from ILI-NET. (B.) Estimated percent vaccine effectiveness of the influenza vaccine per season from MMWR reports. (C.) Mean daily ambient temperature (orange) and barometric pressure (green) data from NOAA. Inclusion of these data sources greatly improves the proposed transmission model forecasts of incident hospitalizations with potential to improve evidence-based public health decision.

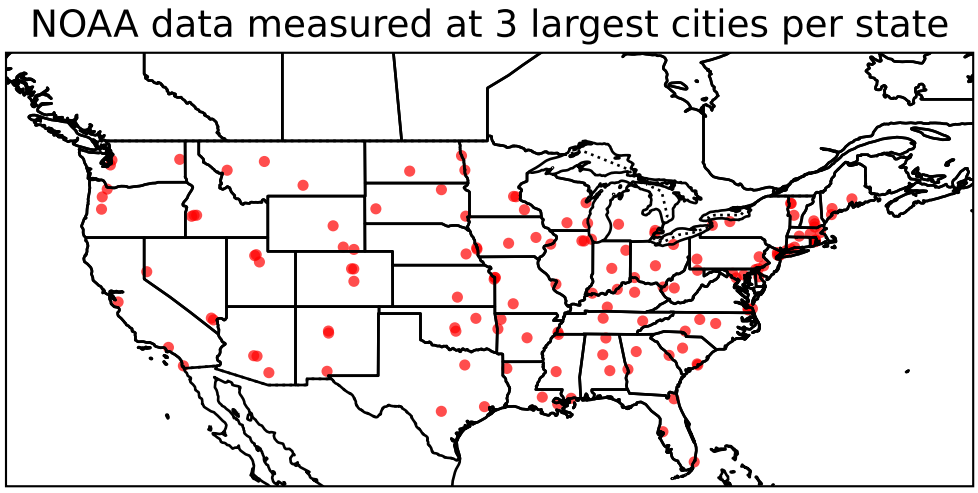

FIG. (2) Ambient temperature and barometric pressure data—known to modulate influenza transmission—was collected from the three most populace cities in every state (red circles show coordinates of cities). The mean daily temperature and atmospheric pressure at the US national level was computed as the average over all city temperatures and pressures. These data were collected from the NOAA Global Historical Climatology Network dataset via the meteostat python package.

II. TRANSMISSION MODEL

Let individuals be assigned to one of 6 disease states: susceptible (S) and not vaccinated, susceptible and vaccinated (S_t), latent (exposed, but not infectious) (E), infectious (I), hospitalized (H), and removed (R). For a more realistic exposed period, split the exposed compartment into a E_1 and E_2 compartment.

Then if we assume a closed population with homogeneous mixing, we expect the proportion of individuals in each disease state to evolve according to Figure 3 where $\beta(t)$ is a time-dependent transmission rate that describes the number of effective contacts (contacts between a susceptible and infected individual that result in influenza transmission); τ is the reduction (a value between zero and one) that describes how vaccination reduces the transmission rate; $1/\sigma$ describes the duration of the latent period (the time between when a susceptible makes effective contact and they become infectious); $1/\gamma$ describes the mean duration of the infectious period; ϕ defines the fraction of individuals who are hospitalized with influenza; and $1/\rho$ describes the average length of time for a hospital stay. [13]

In addition to the above ODE system, we append three 'helper' states that record the number of cumulative incident: infections, reported ILI; and hospitalizations.

$$\dot{cI} = 2\sigma E_2;\ \dot{cILI+} = \alpha\gamma I;\ \dot{cH} = \phi\gamma I$$

From the cumulative incident number ILI+ reports, and hospitalizations we can compute incident ILI+ reports and hospitalizations by taking first differences.

If the number of weeks starts at $t = 1$ and ends at $t = T$ then we define the initial conditions for all eight states (because we include ILI and the split-compartments E_1 and E_2) at time $t = 0$. This is so that after taking first differences we arrive at T incident ILI reports and hospitalizations.

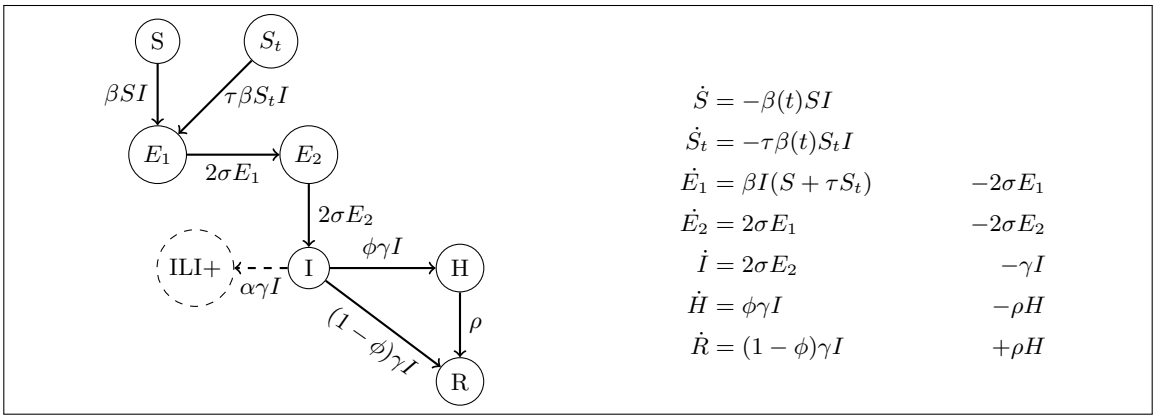

FIG. (3) (Left) A flow diagram that presents how individuals move through disease states in this dynamical system (right). Disease states are represented as circles. Rates are placed on arrows to describe the rate at which individual move from one state to another. The ILI state is placed in a dashed circle because this state is only an observed state that is used to inform prevalent infections (I).

III. INITIAL CONDITIONS

We need to estimate the initial proportions of seven disease states (plus the three helper states) without observing these proportions in our collected dataset. In what follows, when we make reference to a disease state value the reader should assume that this is the *initial value* that begins the system. In other words the initial proportion of infected in this section is denoted I instead of I_0. This is to ease notation.

To reduce the number of initial states to estimate, we assume that $E_1 = E_1 = R = H = cH = cILI+ = 0$. We will assume that there exists a total proportion of susceptibles, $S_{\text{ttl}} = S + S_t$, and that a proportion (v) of these individuals are vaccinated or begin in the S_t compartment. This means that

$$S_t = vS_{\text{ttl}};\ S = (1-v)S_{\text{ttl}}.$$

We placed Normal priors over the proportion of susceptibles S_{ttl} on the logit scale as

$$\text{logit}\left(S_{\text{ttl}}\right) \sim \mathcal{N}\left(\text{logit}\left(\widehat{S_{\text{ttl}}}\right), \sigma_S\right).$$

where $\widehat{S_{ttl}}$ is a point estimate (see below section V titled Maximum likelihood estimates to support Bayesian model). We could have decided to use a Beta density as a prior over the proportion of susceptibles. Our choice to map the proportion of susceptibles to logit-space and use a Normal density is because this improved our fit via variational inference (details on inference can be found below in Section VI titled Bayesian hierarchical model).

We estimate the initial proportion of infections I as one minus the proportion of all susceptibles or

$$I = (1 - S_{\text{ttl}}).$$

We also set the initial proportion of cumulative incident infections, cI, equal to I. Then our final vector of initial conditions is length 10 and equal to

$$(S, S_t, 0, 0, I, 0, 0, I, 0, 0)$$

where we have placed zeros for all states except for the susceptible and infected disease states. The primary reason we chose to set only four out of the ten initial conditions to positive values is to curb model non-identifiability. Several different initial conditions for this dynamical system likely lead to the same log-likelihood value. Rather than set arbitrary values for all ten disease states we attempted to constrain, as much as possible, the range of initial conditions.

IV. FIXING EPIDEMIOLOGICAL PARAMETERS

The above dynamical system can produce the same vector of incident hospitalizations for numerous parameter combinations, an issue called non-identifiability [14, 15]. Our choice of initial conditions is one approach for reducing non-identifiability. An additional method to address non-identifiability is to fix specific parameters in the above dynamical system based on previous literature about influenza dynamics. In support of these fixed values, we find that our estimated reproduction numbers for all locations are in the interval one to two, close to typical values found in past literature (See Figure 5 of reproduction numbers).

Parameter	Fixed value	Literary sources
$1/\gamma$	3/7 week	[16, 17]
$1/\sigma$	2/7 of a week	[17]
$1/\rho$	5/7 of a week	[18, 19]

TABLE (I) To reduce issues of non-identifiability, we fixed the infectious period ($1/\gamma$) to three days; the exposed period ($1/e$) to two days; the hospitalization period ($1/\rho$) to five days; and the initial proportion of vaccinated individuals to fifty percent. Past literature to support these values are provided.

V. MAXIMUM LIKELIHOOD ESTIMATES TO SUPPORT BAYESIAN MODEL

In addition to the above parameters whose values are fixed, the following parameters are fit to ILI+ and incident hospitalization data: S_{ttl}, the total percent of susceptibles (both vaccinated and un-vaccinated); β, the transmission rate; α, the proportion of infected individuals who are reported as an ILI+ case; ϕ the proportion of individuals who are reported as hospitalized due to influenza. The model is fit using a Genetic

algorithm with a population size of 10,000. [20] The loglikelihood that is maximized is defined as

$$\ell\ell(\theta) = \sum_{s=1}^{S}\sum_{t=1}^{T} \log\left[\text{Pois}(h_{s,t}|N\hat{h}_{s,t})\right] + \sum_{s=1}^{S}\sum_{t=1}^{T} \log\left[\text{BetaBinomial}(\text{ILI}_{s,t}|\zeta\widehat{\text{ILI}}_{s,t}, \zeta\left(1-\widehat{\text{ILI}}_{s,t}\right), \text{NILI}_{s,t})\right]$$

where the probability assigned to the observed number of incident hospitalizations $h_{s,t}$ in season s and week t, given the fitted estimate $\hat{h}_{s,t}$ and population size N equals

$$\text{Pois}(h_{s,t}|N\hat{h}_{s,t});$$

and where

$$\text{BetaBinomial}(\text{ILI}+_{s,t}|\zeta\widehat{\text{ILI}+}_{s,t}, \zeta\left(1-\widehat{\text{ILI}+}_{s,t}\right), \text{NILI}+_{s,t})$$

is the probability of observing $\text{ILI}+_{s,t}$ cases in season s and week t out of a total of $\text{NILI}+_{s,t}$ total patients, given the model fitted number of observed ILI+ reports ($\widehat{\text{ILI}+}_{s,t}$). Note the BetaBinomial has a concentration parameter ζ. We set ζ to 100 so the model focuses both on hospitalizations and ILI data.

The model we assume is a simplified version of the full Bayesian model specified below. The maximum likelihood estimates computed from this model will be used below to help set priors and serve as starting points for the fuller Bayesian analysis.

VI. BAYESIAN HIERARCHICAL MODEL

We will describe our Bayesian specification in three parts: (1) how MLE estimates are used to shape posterior parameter estimates; (2) how we build our data-driven time-dependent transmission rate; and (3) a 'discrepancy' term to account for a mis-specified dynamical system. [14, 21]

MLE estimates are used to define priors over the transmission rate, proportion of observed ILI+ reports, proportion of observed hospitalizations, and the initial proportion of susceptible individuals.

$$\begin{aligned}
&\sigma_\beta \sim \text{Half-Cauchy}(10);\ \log(\beta_0) \sim \mathcal{N}(\beta_{0,\text{mle}}, \sigma_\beta)\\
&\sigma_{\beta s} \sim \text{Half-Cauchy}(1); \log(\beta_{0,s}) \sim \mathcal{N}(\log(\beta_0), \sigma_{\beta s})\\
&\sigma_\phi \sim \text{Half-Cauchy}(10);\ \text{logit}(\phi) \sim \mathcal{N}(\phi_{\text{mle}}, \sigma_\phi)\\
&\sigma_\alpha \sim \text{Half-Cauchy}(10);\ \log(\alpha) \sim \mathcal{N}(\alpha_{\text{mle}}, \sigma_\alpha)\\
&\text{logit}(S_{\text{ttl}}) \sim \mathcal{N}(S_{\text{ttl,mle}}, 10)\\
&S_0 = (1-\nu)S_{\text{ttl}};\ S_t = \nu S_{\text{ttl}}\\
&I_0 = (1 - S_{\text{ttl}})
\end{aligned}$$

Parameters that are defined to be positive are sampled on log scale and parameters defined on the unit interval are sampled on the logit scale. Note that the intercept for transmission rates, β_0, is defined hierarchically across seasons. This assumes that the baseline transmission across influenza seasons is similar—though not exactly the same. Below, we add to this baseline transmission rate two season-specific covariates.

To incorporate vaccine effectiveness data from MMWR [3], we assume that the reduction in transmission for vaccinated individuals is equal to the estimated vaccine effectiveness for that season

$$\tau_s = \text{VE}_{\text{MMWR},s}$$

For our (yet unseen to the model) 2023/34 season, we set τ to be the mean of all collected vaccine effectiveness values. This is a simplifying assumption and a more rigorous approach may be to first propose a density over these values from which to to sample.

To incorporate into the model climate data from NOAA, we assume that the transmission rate is a function of temperature, temp, pressure pres as

$$\log(\beta_{s,t}) = \log(\beta_{0,s}) + b_1 \text{temp}_{s,t} + b_2 \text{pres}_{s,t}$$

Note that the term $\log(\beta_{0,s})$ was defined above and links transmission rates across seasons. We did not include an additional 'error' term in the above time-dependent transmission rate in order to emphasize the data-driven component of the model. An error term could be included and may improve model fit.

The above prior densities are used to produce a set of S (one per season) model-proposed trajectories for weekly incident ILI+ reports $\widehat{\text{ILI}}_{s,t}$ and incident hospitalizations $\hat{h}_{s,t}$. The ODE system was integrated by Euler's method with a step size of 1/7 (i.e. on the scale of one day). Integration took place in diffrax. [22]

The current model is similar to a state-space model where the latent, disease states are deterministic. To add noise that should be propagated forward in time, we assume that the number of incident hospitalizations is sampled from a Gaussian Process with covariance function equal to Brownian motion.

$$\begin{aligned}
\text{logit}\left(h*_{s,t}\right) &\sim \text{MVN}(\hat{h}_{s,t}, K(t_1, t_2, \omega)) \\
\mu &\sim \text{Beta}(1,1);\ \eta \sim \text{LogNormal}(0,1) \\
\omega_s &\sim \text{LogNormal}(\mu, \eta) \\
K_s(t_1, t_2) &= \omega_s \min(t_1, t_2)
\end{aligned}$$

where MVN is a Multivariate Gaussian Density. Note that ω, the dispersion for Brownian motion, is defined hierarchically over seasons. Then the log likelihood is calibrated to both incident hospitalization data and ILI+ data:

$$\zeta \sim \text{Half-Cauchy}(1)$$

$$\ell\ell(\theta) = \sum_{s=1}^{S}\sum_{t=1}^{T} \log\left[\text{Negative Binomial}(h_{s,t}|Nh*_{s,t}, 50)\right] +$$

$$\sum_{s=1}^{S}\sum_{t=1}^{T} \log\left[\text{BetaBinomial}(\text{ILI}_{s,t}|\zeta \hat{\text{ILI}}_{s,t}, \zeta\left(1 - \hat{\text{ILI}}_{s,t}\right), \text{NILI}_{s,t})\right]$$

Rather than use the stricter Poisson density for incident hospitalizations, we chose a Negative Binomial density with mean $Nh*_{s,t}$ and concentration 50. [23] The Negative Binomial with an infinite concentration is equivalent to a Poisson density, and so our choice here allows for over-dispersion in the collected hospitalization data.

The fully specified Bayesian model is fit using stochastic Variational inference. [24] Numpyro was used to implement this model. [25] We used a Normal density with diagonal covariance matrix as the 'guide' or approximate density. We chose Adam as the optimizer with a step size of 0.001 and ran the VI algorithm 2×10^4 iterations. This implementation led to a decreasing ELBO and very little change after 2×10^4 iterations.

VII. FORECAST

A forecast is generated by sampling from the estimated posterior density that was computed with Variational inference. This sampling procedure does not take into account variation in the, yet unseen, covariate data. Instead we assume that the covariate data for the upcoming data is the average over past seasons—a simplifying assumption. Forecasts from 1-32 weeks ahead are generated and 23 quantiles are computed that are the same quantiles as requested by the CDC FluSight challenge. [26].

VIII. FORECAST EVALUATION

As a formal evaluation, we compared the weighted interval score (WIS) and the absolute error (AE) between the median forecast and true number of US national incident hospitalizations for all forecast horizons (See Figure 4). For both the WIS and AE, smaller values indicate a better performing model [27]. For all forecast horizons, and with both evaluation metrics, the transmission model trained on all data sources (Fig. 4 blue) outperforms the model trained only on NHSN data (Fig. 4 red).

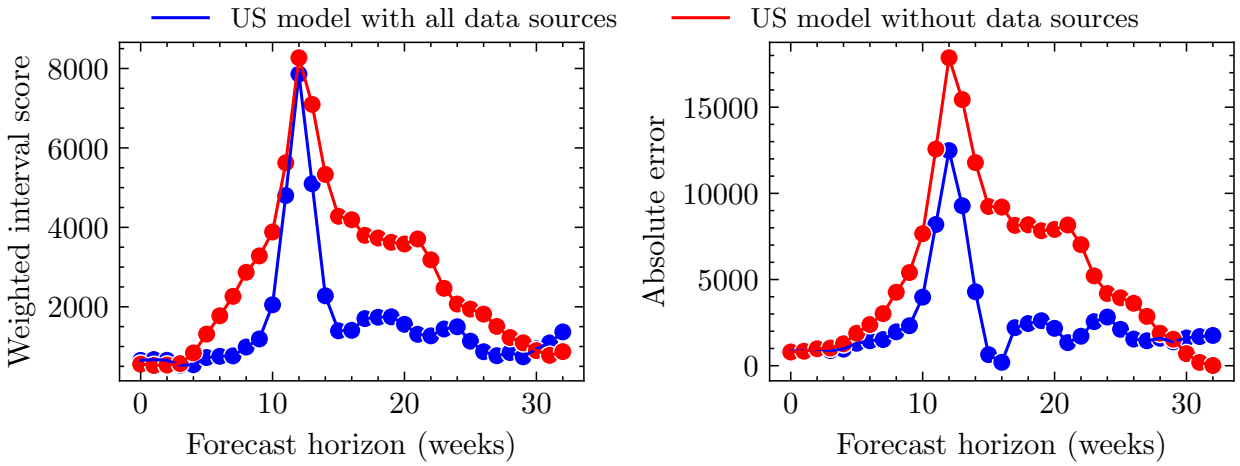

FIG. (4) A comparison of (Left) the weighted interval score and (Right) the absolute error between the median forecast and truth for 1-32 week ahead forecasts of US national incident hospitalizations for two models: (blue) the model trained with all data sources and (red) the model trained with only NHSN data.

IX. TIME-DEPENDENT EFFECTIVE TRANSMISSION RATE

The time-dependent transmission rate depends on the average temperature and pressure data from NOAA (See Figure 5). From the transmission rate we can compute the effective reproduction number as

$$\mathcal{R}_{eff} = \beta(t)(1+\tau)\left(\frac{1}{\gamma}\right)$$

As expected, for the red model that assumes a constant transmission, the average effective reproduction number is 1.6—consistent with past influenza modeling efforts. [28] For the blue model, the effective reproduction number is highest at the beginning of the season. Then the rate drops, still within reasonable estimates for influenza, to values as low as 1.2. The uncertainty around this value is small, and should be corrected with a more advanced model.

X. DATA AVAILABILITY, MODEL CODE, AND REPRODUCIBLE PIPELINE

The data, model code are available at GitHub at https://github.com/computationalUncertaintyLab/importance_of_data. A Makefile can be used to run the code and produce model data and all the figures presented in the main manuscript. This makes clear what data was used, how it was used, and how the data

was mapped to a US national forecast for the entire 2023/24 season. Comments or questions on the code can be either emailed to the corresponding author or submitted directly on the above GitHub link.

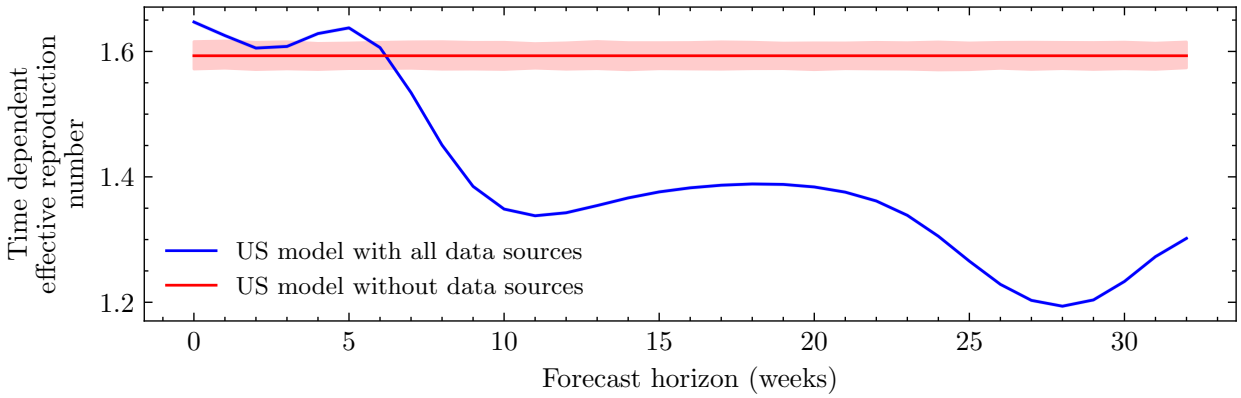

FIG. (5) Mean and 95% uncertainty interval for the time-dependent transmission rate $\beta(t)$ for the model trained on all data sources (blue) and model trained only on NHSN (red). The time dependent transmission is defined over all 32 weeks and depends on the average temperature and pressure data from NOAA (via meteostat).

[1] Health and Human Services. Weekly hospital respiratory data, 2025.
[2] Centers for Disease Control and Prevention. Outpatient respiratory illness activity map determined by data reported to ilinet, 2025.
[3] Centers for Disease Control and Prevention. Morbidity and mortality weekly report, 2025.
[4] National Centers for Environmental Information. Global historical climatology network daily (ghcnd), 2025.
[5] Craig J McGowan, Matthew Biggerstaff, Michael Johansson, Karyn M Apfeldorf, Michal Ben-Nun, Logan Brooks, Matteo Convertino, Madhav Erraguntla, David C Farrow, John Freeze, et al. Collaborative efforts to forecast seasonal influenza in the united states, 2015–2016. *Scientific reports*, 9(1):683, 2019.
[6] Michal Ben-Nun, Pete Riley, James Turtle, David P Bacon, and Steven Riley. Forecasting national and regional influenza-like illness for the usa. *PLoS computational biology*, 15(5):e1007013, 2019.
[7] Sen Pei and Jeffrey Shaman. Aggregating forecasts of multiple respiratory pathogens supports more accurate forecasting of influenza-like illness. *PLoS computational biology*, 16(10):e1008301, 2020.
[8] Nicholas G Reich, Logan C Brooks, Spencer J Fox, Sasikiran Kandula, Craig J McGowan, Evan Moore, Dave Osthus, Evan L Ray, Abhinav Tushar, Teresa K Yamana, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019.
[9] Sen Pei, Sasikiran Kandula, Wan Yang, and Jeffrey Shaman. Forecasting the spatial transmission of influenza in the united states. *Proceedings of the National Academy of Sciences*, 115(11):2752–2757, 2018.
[10] Anice C Lowen, Samira Mubareka, John Steel, and Peter Palese. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS pathogens*, 3(10):e151, 2007.
[11] J Leith Holloway Jr. Smoothing and filtering of time series and space fields. In *Advances in geophysics*, volume 4, pages 351–389. Elsevier, 1958.
[12] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.
[13] Fred Brauer, Carlos Castillo-Chavez, Zhilan Feng, Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng. Models for influenza. *Mathematical Models in Epidemiology*, pages 311–350, 2019.
[14] Dave Osthus, James Gattiker, Reid Priedhorsky, and Sara Y Del Valle. Dynamic bayesian influenza forecasting in the united states with hierarchical discrepancy (with discussion). 2019.
[15] Dave Osthus, Kyle S Hickmann, Petruţa C Caragea, Dave Higdon, and Sara Y Del Valle. Forecasting seasonal influenza with a state-space sir model. *The annals of applied statistics*, 11(1):202, 2017.
[16] Benjamin J Cowling, Kwok Hung Chan, Vicky J Fang, Lincoln LH Lau, Hau Chi So, Rita OP Fung, Edward SK Ma, Alfred SK Kwong, Chi-Wai Chan, Wendy WS Tsui, et al. Comparative epidemiology of pandemic and seasonal influenza a in households. *New England journal of medicine*, 362(23):2175–2184, 2010.
[17] Mohsen Moghadami. A narrative review of influenza: a seasonal and pandemic disease. *Iranian journal of medical sciences*, 42(1):2, 2017.
[18] Lin Dou, Dan Reynolds, Lindsey Wallace, John O'Horo, Rahul Kashyap, Ognjen Gajic, and Hemang Yadav. Decreased hospital length of stay with early administration of oseltamivir in patients hospitalized with influenza. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, 4(2):176–182, 2020.
[19] Nelson Lee, Paul KS Chan, Kin Wing Choi, Grace Lui, Bonnie Wong, Clive S Cockram, David SC Hui, Raymond Lai, Julian W Tang, and Joseph JY Sung. Factors associated with early hospital discharge of adult influenza patients. *Antiviral therapy*, 12(4):501–508, 2007.
[20] Julian Blank and Kalyanmoy Deb. Pymoo: Multi-objective optimization in python. *Ieee access*, 8:89497–89509, 2020.
[21] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.
[22] Patrick Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021.
[23] Numpyro. Negative binomial density, 2025. Accessed: 2025-01-31.
[24] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.
[25] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.
[26] Sarabeth M Mathis, Alexander E Webber, Tomás M León, Erin L Murray, Monica Sun, Lauren A White, Logan C Brooks, Alden Green, Addison J Hu, Roni Rosenfeld, et al. Evaluation of flusight influenza forecasting in the 2021–22 and 2022–23 seasons with a new target laboratory-confirmed influenza hospitalizations. *Nature*

communications, 15(1):6289, 2024.

[27] Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2):e1008618, 2021.

[28] Matthew Biggerstaff, Simon Cauchemez, Carrie Reed, Manoj Gambhir, and Lyn Finelli. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC infectious diseases*, 14(1):1–20, 2014.