Copyright

by

Kai Liu

2019

The Dissertation Committee for Kai Liu Certifies that this is the approved version of the following Dissertation:

Improving Surveillance and Prediction of Emerging and Re-emerging Infectious Diseases

Committee:

Lauren Ancel Meyers, Supervisor

Claus O. Wilke

James J. Bull

David Hillis

James Scott

Improving Surveillance and Prediction of Emerging and Re-emerging Infectious Diseases

by

Kai Liu

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas at Austin

August 2019

To everyone in my life

Do not wait;

the time will never be ‘just right’.

Start where you stand,

and work with whatever tools you may have at your command,

and better tools will be found as you go along.

GEORGE HERBERT

Acknowledgements

It has been over two thousand days since I started my Ph.D. journey. I still remember that a former coordinator in my Ph.D. program told us in the first day that only about half of the students in the program could finish the degree. There are multiple barriers you have to pass during the years, like passing all core courses, passing preliminary exam, passing qualify exam *etc*.. Not to mention, you need to find a topic that no one had never worked on, and put all your efforts on it over the next five or six years. I have never imagined that one day I am able to get here. The whole Ph.D. journey would not have happened without the support of everyone I have met, not only those over the past six years, but everyone in my life.

To my parents, I cannot imagine how much efforts are required to raise a child, especially the only child in the family. I am like the king of the family. I cannot forget how early you had to get up in the morning to prepare a big breakfast for me every day during my three-year high school life. I cannot forget how late you companied with me every night when I was working on my homework. I cannot forget how generous you are when buying anything for me. I cannot forget how supportive you are when I decided to go to the United States for my Ph.D. study even though you are not willing me to leave the family. I cannot forget how eager you are to have video chat with me every week over the past six years. I was only able to go back to China twice during the past six years. Every time I was at home, you did everything you can to make me enjoy the stay. Even though you do not understand what I was working on in my Ph.D. study, and even may have no idea about graduate school, you did you best to support me, to comfort me when I was under pressure. You always say "No worries. If you cannot make it, just come back home. We still have the ability to raise you." This made me tear up almost every time. I am not able to make this Ph.D. a reality without you.

I would not be able to finish the dissertation without the generous support from my Ph.D. advisor Dr. Lauren Ancel Meyers. Lauren is a such supportive, patient and inspiring person. It is a great honor to work with her over the past five years. I was literally a newbie to statistical modeling and programming when I joined the lab. Lauren was so patient to guide me and give me the time and space to grow up. She also gave me the opportunity and freedom to explore different research topics, and guided me throughout the process. Yes, I admitted how hard it was to find a meaningful topic and organize my ideas at first. But it is the only way to grow up as a scientist. See I made it after six years. I could not forget how many times I messed up over the past six years, but Lauren was always there to help me get through it. Sometimes, I even doubt if I was able to be so patient if I was her. Besides, she is so smart and possesses outstanding communication skills. She can always express her ideas clearly and understand my ideas quickly. Lauren will always be a role model for me, not only in science but also in personality.

I would like to thank my master advisor Dr. Jing He as well. Without her, I was not even able to sit here. I still remember a meeting with her eight years ago when I just started my second year in her lab. I discussed my career plan after graduation with her. I planned to go to Germany for doctoral study, while she suggested me to go to the United States. I told her this was too tough to me since I have to take both TOEFL and GRE examinations and plus it takes about five years to finish the Ph.D. in the US. She laughed: “Come on! Are you worried about your age? That does not really matter. Just do that, and you have my fully support.” Yes, she did do what she said. When I asked for time off to prepare those examinations, she said yes. When I needed recommendation letters for over ten universities, she did every one without complains. When I was waiting for application results, she did her best to recommend me to her collaborators in the US.

I would also like to say thank you to a number of people without whom I may not finish my research work within six years. They are my committee members: Dr. James Scott, Dr. Claus Wilke, Dr. David Hillis, and Dr. James Bull. Thank you for your feedback and suggestions on my research every time we met. They are my collaborators, past and current Meyers lab members: Ravi Srinivasan, Zhanwei Du, Joel Miller, Spencer Fox, Lauren Castro, Zeynep Ertem, Jose Luis Herrera Diestra, Steve Bellan, Ned Dmitrov, Amanda Perofsky, Xutong Wang, Xi Chen, Remy Pasco, and Briana Betke. You guys always provide me informative feedbacks on my research via lab meeting or emails. Thank you for all your time and efforts.

Graduate school is not only about research, but also about life. I am so grateful to have some interesting and cheerful friends here in Austin. We always gather together in free time to find good food, watch movies, go to concerts, do some fun stuffs like the

Wipeout Run, the Color Run, kayaking *etc.*. We even have online chatting groups to complain the life in graduate school (even though sometimes it was not that bad), research difficulties, sometimes even just bullshitting. We had so many golden memories here in Austin. I am not sure if I could get through graduate school without you guys.

I would also like to acknowledge various groups at the University of Texas at Austin: the Graduate School, the International Office, my home institution – the Institute for Cellular and Molecular Biology, the Department of Integrative Biology, Texas Advanced Computing Center. They helped me handle various processes and documentations throughout my Ph.D. study. Without their support, I was not able to focus on my research and be productive.

My acknowledgement is not complete without mentioning funding support from the following sources: a US Department of Defense the Defense Threat Reduction Agency contract HDTRA-14-C-0114 to LA Meyers, S Barber and A Ellington, a US National Institute of General Medical Sciences Models of Infectious Disease Agent Study Grant U01GM087719 to LA Meyers and AP Galvani, a Center for Disease Control and Prevention contract to LA Meyers, and a UT Graduate School travel award to K Liu. With these funding, I was able to do research without worries, travel to different places for conferences and meet scientists from different organizations, and most importantly, feed myself over the past six years. It really helped me reshape my outlook on the world.

Abstract

Improving Surveillance and Prediction of Emerging and Re-emerging Infectious Diseases

Kai Liu, Ph.D.
The University of Texas at Austin, 2019

Supervisor: Lauren Ancel Meyers

Infectious diseases are emerging at an unprecedent rate in recent years, such as the flu pandemic initialized from Mexico in 2009, the 2014 Ebola epidemic in West Africa, and the 2016-2017 expansion of Zika across Americas. They rarely happened previously and thus lack resources and data to detect and predict their spread. This highlights the challenges in emerging an re-emerging infectious disease surveillance. In the dissertation, I mainly put efforts in developing methods for early detection of such diseases, and assessing predictive power of various models in early phase of an epidemic. In Chapter 2, I developed a two-layer early detection framework which provides early warning of emerging epidemics based on the idea of anomaly detection. The framework could evaluate and identify data sources to achieve the best performance automatically from available data, such as data from the Internet and public health surveillance systems. I demonstrated the framework using historical influenza data in the US, and found that the optimal combination of predictors includes data sources from Google search query and Wikipedia page view. The optimized system is able to detect the onset of seasonal influenza outbreaks an average of 16.4 weeks in advance, and the second wave of the 2009 flu pandemic 5 weeks ahead. In Chapter 3, I extended the framework in Chapter 2 to identify large dengue outbreaks from small ones. The results show that the framework could personalize optimal combinations of predictors for different locations, and an optimal combination for one location might not perform well for other locations. In Chapter 4, I investigated the contribution of different population structures to total epidemic incidence, peak intensity and timing, and also explored the ability of various models with different population structures in predicting epidemic dynamics. The results suggest that heterogeneous contact pattern and direct contacts dominate the evolution of epidemics, and a homogeneous model is not able to provide reliable prediction for an epidemic. In summary, my dissertation not only provides method frameworks for building early detection systems for emerging and re-emerging infectious diseases, but also gives insight to the effects of various models in predicting epidemics.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The development of machine learning and big data makes infectious disease forecasting in real time possible. Center for Disease Control and Prevention (CDC) started to host 'Predict the Influenza Season Challenge' in 2013 to solicit prospective, real-time weekly forecasts of regional weighted influenza-like illness (wILI) measures from teams across the world [1,2]. In collaboration with some US federal government agencies and state public health officials, CDC also organizes a platform *Epidemic Prediction Initiative* [3] to host infectious diseases forecasting challenges regularly. The purpose of the platform is to (1) share epidemiological data with research communities; (2) develop metrics relevant to decision-maker for evaluating forecasting models; (3) move forecasting from research to public health decision-making. It has drawn a lot of attention of research communities, and many forecasting methods have been developed. For example, Shaman *et al.* developed a humidity-forced susceptible-infectious-recovered-susceptible (SIRS) ensemble adjustment Kalman filter (EAKF) forecasting framework by adapting ideas from weather forecasting [4,5]. Brooks *et al.* developed an empirical Bayes framework by constructing prior distributions based on historical data [6]. Hickmann *et al.* combined data assimilation methods with Wikipedia page view data relevance to influenza and CDC wILI reports to create weekly forecast for seasonal influenza [7]. However, forecasts made by these models rely heavily on historical data from which to learn historical patterns. This is not possible to be obtained for emerging or re-emerging infectious diseases. For example, the pandemic flu originating from Mexico in 2009 started in April which is not consistent with seasonal influenza (usually in winter) [8,9], and thus forecasting models for seasonal influenza are not able to capture it. Consequently, when training forecasting models for seasonal influenza, the pandemic flu-affected influenza season is usually discarded [5–7]. The 2014-2015 Ebola epidemic occurred in West Africa was the largest one since it appeared [10]. It is impossible to build forecasting model for the Ebola epidemic. These threads highlight the critical shortcomings in infectious diseases surveillance and the urgent needs for building models to detect and predict emerging and re-emerging infectious diseases. With reliable surveillance systems, public health officials are able to allocate resources, plan hospital bed capacity, and distribute antivirals and vaccines in advance.

My dissertation focuses on two aspects in building surveillance systems for emerging and reemerging infectious diseases: (1) develop method for detecting emerging and re-emerging outbreaks as early and accurately as possible. (2) assess the tradeoff between model complexity and prediction reliability.

In Chapter 2, I developed a hierarchical framework to build early detection systems for infectious diseases. The framework couples a Multivariate Exponentially Weighted Moving Average (MEWMA) model with a forward feature selection (FFS) algorithm. The MEWMA model is adapted from an anomaly detection method, which assumes that observations below an event threshold follow a multivariate Gaussian distribution (*null* distribution). If an incoming observation is away from the null distribution, an alarm will be triggered. The FFS algorithm in the framework is used to evaluate thousands of candidate data sources sequentially and identify small combinations maximizing the performance of the MEWMA model. The framework is demonstrated using historical influenza data in the US from 2009-2017. We found that the optimal combination of data sources for early detection of influenza outbreaks in the US includes six Google search time series and two Wikipedia page view time series. With the optimal combination of data sources, the system is able to sound alarms for the onset of seasonal influenza an average of 16.4 weeks prior to the CDC-defined 2% threshold. Moreover, the system also triggers an alarm for the second wave of the 2009 flu pandemic five weeks in advance, which outperforms baseline models. The MEWMA-FFS framework, which can be applied to any infectious diseases with any number of candidate data sources, has been implemented as a user-friendly app in the Biosurveillance Ecosystem (BSVE) build by the US Defense Threat Reduction Agency (DTRA).

Dengue has been endemic in populations across many tropical and sub-tropical countries. By plotting historical dengue incidence data, I found that dengue virus usually causes large outbreaks in those countries during some but not all years, and no explicit seasonal pattern is observed between large outbreaks. This indicates a scenario where our MEWMA-FFS framework can be useful. Therefore, in Chapter 3, I adapted the MEWMA-FFS framework for detecting large dengue waves by making three modifications: (1) introduce a parameter *baseline threshold* to the MEWMA model so that it has the power to identify large outbreaks from small ones; (2) integrate a penalty term to false alarms into the objective function. Consequently, the number of false alarms can be controlled by adjusting the penalty term. It is faster and more intuitive than the Average Time between False Signals (ATFS)-based method in Chapter 2; (3) the simulation-based parameter optimization method is replaced by a Bayesian-based algorithm. It decreases the required computing time from 144 hours to 48 hours when over 100 candidate data source are being optimized on the cluster Olympus [11]. I applied the modified framework to optimize predictor data sources for detecting large dengue outbreaks in three study areas: the Country of Mexico, San Juan metropolitan area in Puerto Rico, and Iquitos metropolitan area in Peru. The results show that the framework is sensitive to locations—different data sources are selected as optimal predictors for different locations, and optimized predictor data sources for one location are not informative for other locations.

As an epidemic is emerging, epidemiologists usually use mathematical models to predict future trajectory of an epidemic. The prediction can help public health officials with resources allocation, and intervention implementation. A critical question in this field is how complex a model should be for making reliable predictions. During the 2014 Ebola epidemic in West Africa, predictive models without considering population structures overestimated the total incidence significantly. Even though the overestimation might stem from effective intervention strategies, population structures are also potential factors causing the bias. In Chapter 4, I derived ordinary differential equations to model infectious disease transmission on contact networks with various population structures, including heterogeneous contact pattern, directed contacts and clustering. In an ideal scenario where the contact network and disease transmission parameters are known, I explored the contributions of different structures to total incidence, peak intensity and timing. I found that heterogeneous and directed contacts are dominate factors in driving epidemic dynamics, while the effect of clustering is minor. Using data collected in early phase of an simulated epidemic, we further investigated the ability of various models to infer transmission rates and make predictions based on estimated transmission rates. This is similar to the workflow of epidemic prediction in practice. I found that a model ignoring all three population structures always overestimate the total incidence, peak timing and intensity by more than 10%, 20%, and about 6 days, while a model considering only heterogeneous contact pattern is able to improve the prediction by 5%, 20% and 3 days.

Chapter 2: Early detection of influenza outbreaks in the United States[1]

2.1 ABSTRACT

Public health surveillance systems often fail to detect emerging infectious diseases, particularly in resource limited settings. By integrating relevant clinical and internet-source data, we can close critical gaps in coverage and accelerate outbreak detection. Here, we present a multivariate algorithm that uses freely available online data to provide early warning of emerging influenza epidemics in the US. We evaluated 240 candidate predictors and found that the most predictive combination does *not* include surveillance or electronic health records data, but instead consists of eight Google search and Wikipedia pageview time series reflecting changing levels of interest in influenza-related topics. In cross validation on 2010-2016 data, this model sounds alarms an average of 16.4 weeks prior to influenza activity reaching the Center for Disease Control and Prevention (CDC) threshold for declaring the start of the season. In an out-of-sample test on data from the rapidly-emerging fall wave of the 2009 H1N1 pandemic, it recognized the threat five weeks in advance of this surveillance threshold. Simpler models, including fixed week-of-the-year triggers, lag the optimized alarms by only a few weeks when detecting seasonal influenza, but fail to provide early warning in the 2009

[1] Considerable portions of this chapter were preprinted on arXiv as Liu, K, Srinivasan, R, and Meyers, LA. 2019. Early detection of influenza outbreaks in the United States. arXiv:1903.01048v1. **Contributions:** KL, RS, LAM designed research; KL and RS performed research; KL and LAM analyzed data; KL, RS, and LAM wrote the paper.

pandemic scenario. This demonstrates a robust method for designing next generation outbreak detection systems. By combining scan statistics with machine learning, it identifies tractable combinations of data sources (from among thousands of candidates) that can provide early warning of emerging infectious disease threats worldwide.

2.2 INTRODUCTION

Emerging and re-emerging human viruses threaten global health and security. Early warning is vital to preventing and containing outbreaks. However, viruses often emerge unexpectedly in populations that lack resources to detect and control their spread. The silent Mexican origin of the 2009 pandemic [8,9], unprecedented 2014-2015 expansion of Ebola out of Guinea [10], and the rapid spread of Zika throughout the Americas in 2016-2017 [12] highlighted critical shortcomings and the potential for life-saving improvements in global disease surveillance.

Traditionally, public health agencies have relied on slow, sparse and biased data extracted during local outbreak responses or collected via voluntarily reporting by healthcare providers. The 21st century explosion of health-related internet data--for example, disease-related Google searches, Tweets, and Wikipedia term visits--and the proliferation of pathogen molecular data and electronic health records have introduced a diversity of real-time, high-dimensional, and inexpensive data sources that may ultimately be integrated into or even replace traditional surveillance systems. In building 'nextgen' surveillance systems, we face the interdependent challenges of identifying combinations of data sources that can improve early warning and developing powerful statistical methods to fully exploit them.

Engineers have designed anomaly detection methods for statistical process control (SPC)---including the Shewhart [13], cumulative sum (CUSUM) [14,15], and exponential weighted moving average (EWMA) methods [16]---to achieve real-time detection of small but meaningful deviations in manufacturing processes from single or multiple input data streams. When the focal process is *in-control*, these methods assume that the inputs are independent and identically distributed random variables with distributions that can be estimated from historical data. Anomalous events can thus be detected by scanning real-time data for gross deviations from these baseline distributions.

Biosurveillance systems similarly seek to detect changes in the incidence of an event (e.g., infections) as early and accurately as possible, often based on case *count* data. By adjusting SPC methods to account for autocorrelations, researchers have developed algorithms that can detect the emergence or re-emergence of infectious diseases [17]. Such methods have been applied to influenza [18–22], Ross River disease [23,24], hand-foot-and-mouth disease [25–27], respiratory tract infections [19,28,29], meningitis [30], and tuberculosis outbreaks [31]. These models exploit a variety of public health data sources, including syndromic surveillance, case count and laboratory test data. While they achieve high sensitivity and precision, alarms typically sound once an outbreak has begun to grow exponentially and thus do not provide ample early warning. For annual influenza, CUSUM-derived detection methods applied to Google Flu Trends data sound alarms an average of two weeks prior to the official start of the influenza season [32].

The Early Aberration Reporting System (EARS) [33] was launched by the CDC in 2000s to provide national, state, and local health departments with several CUSUM-derived methods to facilitate the syndromic surveillance. The BioSense surveillance system [34] implements methods derived from EARS to achieve early detection of possible biologic terrorism attacks and other events of public health concern on a national level. Two other surveillance systems, ESSENCE and NYCDOHMH [35,36], maintained by United States Department of Defense and the New York City Department of Health and Mental Hygiene, respectively, implement EWMA-based methods for outbreaks monitoring. Most of these systems are univariate (i.e., analyze a single input data source) and consider only public health surveillance data collected during local outbreak responses or via voluntarily reporting by healthcare providers. The time lag between infection and reporting can be days to weeks. Thus, the earliest warning possible for an emerging outbreak may be well after cases begin rising.

Over the last decade, public health agencies and researchers have begun to explore a variety of 'nextgen' disease-related data sources that might improve the spatiotemporal resolution of surveillance. Electronic health records (EHR) systems like athenahealth can provide near real-time access to millions of patient records, nationally, and have been shown to correlate strongly with influenza activity [37]. Participatory surveillance systems like Flu Near You, which asks volunteers to submit brief weekly health reports, also provide a near real-time view of ILI activity [38]. However, such data sources may be geographically, demographically or socioeconomically biased, depending on the profiles of participating healthcare facilities or volunteers [39]. Internet-source data such as Google Trends [40], Wikipedia page views [7,41], and Twitter feeds [42] exhibit correlations with disease prevalence, and have been harnessed for seasonal influenza nowcasting and forecasting. However, they have not yet been fully evaluated for early outbreak detection, and may be sensitive to sociological perturbations, including media events and behavioral contagion [43,44].

Here, we introduce a hierarchical method for building early and accurate outbreak warning systems that couples a multivariate version of EWMA model with a forward feature selection algorithm (MEWMA-FFS). The method can evaluate thousands of data sources and identify small combinations that maximize the timeliness and sensitivity of alarms while achieving a given level of precision. It can be applied to any infectious disease threat provided sufficient data for the candidate predictors. For novel threats, the candidates may include a wide variety of proxies that are expected to produce dynamics resembling the focal threat (e.g., data on closely related pathogens, other geographic regions, or even social responses to non-disease events).

To demonstrate the approach, we design a multivariate early warning system for influenza outbreaks using eight years of historical data (2009-2017) and hundreds of predictors, including traditional surveillance, internet-source, and EHR data. The optimal combination of input data includes six Google and two Wikipedia time series reflecting online searches for information relating to the symptoms, biology and treatment of influenza. By monitoring these data, the model is expected to detect the emergence of seasonal influenza an average of 16.4 weeks (and standard deviation of 3.3 weeks) in advance of the Center for Disease Control and Prevention (CDC) threshold for the onset of the season. In out-of-sample validation, the model detected the fall wave of the 2009 H1N1 pandemic and the 2016-2017 influenza season five and fourteen weeks prior to this threshold, respectively.

2.3 MATERIALS AND METHODS

2.3.1 Early detection model

The MEWMA model is derived from a method described in [45]. We define one time series as *gold standard*, and one value in the range of the gold standard as the event threshold. Events (outbreaks) correspond to periods when observations in the gold standard cross and remain above the event threshold. We project the timing of events in the gold standard time series onto the candidate time series (predictors). We assume that the data falling outside the event periods follow a multivariate normal distribution $\boldsymbol{F}$ (null distribution) with a mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ that can be estimated from baseline (non-outbreak) data with Equations (2.1) and (2.2):

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X_T} | y_{\boldsymbol{T}} < \varepsilon) \tag{2.1}$$

$$\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X_T} | y_{\boldsymbol{T}} < \varepsilon) \tag{2.2}$$

Here, ε is the value of the threshold defining outbreak events. $\boldsymbol{T}$ are all time points at which observations in gold standard y are below event threshold ε. $\boldsymbol{X_T}$ is a matrix of observations from candidate time series at time points $\boldsymbol{T}$.

At each time t, MEWMA calculates

$$\boldsymbol{S}_t = \begin{cases} \max[\boldsymbol{0}, \lambda(\boldsymbol{X}_t - \boldsymbol{\mu}) + (1-\lambda)\boldsymbol{S}_{t-1}], & \text{for } t > 0 \\ \boldsymbol{0}, & \text{for } t = 0 \end{cases} \tag{2.3}$$

where $\boldsymbol{X}_t$ is a vector of current observation from candidate time series; λ is the smoothing parameter ($0 < \lambda < 1$); $\boldsymbol{S}_t$ is a weighted average of the current observation standardized around $\boldsymbol{\mu}$ and the previous $\boldsymbol{S}$ statistic. Then the multivariate EWMA test statistic E_t is calculated as

$$E_t = \boldsymbol{S}_t^T \boldsymbol{\Sigma}_{\boldsymbol{S}_\infty}^{-1} \boldsymbol{S}_t \tag{2.4}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{S}_\infty} = \frac{\lambda}{2-\lambda} \boldsymbol{\Sigma} \tag{2.5}$$

The MEWMA signals whenever E_t exceeds a predetermined threshold h. That is, the observation at time t deviates significantly from the baseline distribution.

2.3.2 Performance measurement

Given that our objective is to detect emerging outbreaks early and accurately, we evaluate data based on the timing of alarms relative to the start of events. Only alarms within detection windows are considered as true positive alarms. Specifically, we calculate performance of a candidate system (combination of predictors) as given by

$$P(\boldsymbol{X}, \lambda, h; y) = \frac{1}{N}\sum_{n=1}^{N}(1 - \frac{\Delta T_n}{T_w}) \tag{2.6}$$

where N is the total number of events in gold standard, T_w is the length of the detection window (e.g., sixteen weeks surrounding the start of an event) and ΔT_n is the time between the start of the detection window and the first alarm for event n. If no alarm sounds during the detection window for event n, then $\Delta T_n = T_w$. Performance values range from zero to one. A perfect score of one indicates that alarms consistently sound during the first week of the detection window; 0.5 indicates that alarms occur, on average, right at the start of events; lower values indicate delayed alarms, triggered weeks after the event has begun.

Since we do not reset $\boldsymbol{S}_t$ to zero following alarms, the model tends to signal repeatedly until the observations return to baseline. Therefore, we track only the timing of the first alarm during continuous clusters of alarms. MEWMA without resetting saves on computation during data optimization (see Forward feature selection section), as it allows us to reference a single set of stored null distribution calculations when testing for alarms. That is, if $\boldsymbol{F}$ is the null distribution for all candidate time series, we can compute and save the mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{S}_t$ statistic with $\boldsymbol{X}_t$, the vector of observations from all candidate time series at time t. Given a subset $\boldsymbol{U}$ of candidate time series, the test statistic E_t can be computed by using the pre-computed $\boldsymbol{S}_t$ and $\boldsymbol{\Sigma}$ directly.

2.3.3 Parameter optimization

When implementing MEWMA-FFS, we must estimate the smoothing parameter λ and the threshold h. The parameter pair (λ, h) should maximize the performance of the model while minimizing the number of false positive alarms triggered outside detection windows for actual events.

To constrain the number of false positive alarms, we specify the Average Time between False Signals (ATFS) during the training process. This parameter is the expected number of time steps between signals during non-outbreak periods and is given by

$$ATFS \triangleq \mathbb{E}(t^{**} - t^{*} | \tau_{\boldsymbol{s}} = \infty) \tag{2.7}$$

where t^{*} denotes the time an initial alarm is triggered; t^{**} is the next time an alarm sounds; $\tau_{\boldsymbol{s}}$ is the first day of an event, with $\tau_{\boldsymbol{s}} = \infty$ indicating that an event never occurs. The value of ATFS can be estimated using simulations. We first generate samples from the null distribution (data outside event periods), then use the MEWMA procedure described in Equations (2.1) - (2.5) to trigger alarms, and finally use the spacing between these false alarms to estimate ATFS [17].

To calculate the optimal parameter pair, we begin with fixing a value of ATFS (φ). Given a set of time series $\boldsymbol{X}$, this constrains the possible choices for parameter pairs (λ, h) to a curve $\Gamma(\varphi, \boldsymbol{X})$. The overarching optimization goal is given by

$$\boldsymbol{X}^*, \lambda^*, h^* = \arg \max_{\{\boldsymbol{X} \subset \boldsymbol{\Omega} : |\boldsymbol{X}| = k, (\lambda, h) \in \Gamma(\varphi, \boldsymbol{X})\}} P(\boldsymbol{X}, \lambda\,, h; y) \tag{2.8}$$

where $\boldsymbol{X}^*$ is the optimal combination of time series; $\boldsymbol{\Omega}$ is a set of all candidate time series; k is the pre-determined number of time series in the optimization; λ^* and h^* are the optimal parameter pair.

To evaluate parameter pairs (λ, h) on the curve $\Gamma(\varphi, \boldsymbol{X})$, we consider values of λ between zero and one with a step size 0.1. Since ATFS is monotonically increasing in h, this allows us to efficiently find the corresponding approximate value of h using the secant method [46] with the tolerance value of 0.5 and the maximum number of iterations of 100. We plug each resulting parameter pair into the MEWMA model and measure in-sample performance. The parameter pair maximizing the in-sample performance is chosen for out-of-sample prediction.

2.3.4 Forward feature selection

To choose the optimal combinations of time series for early warning, we implement stepwise forward feature selection algorithm in combination with MEWMA. We begin with no predictors and test the model performance (in terms of the average timing of early detection) when we add each of the possible candidate predictors on their own. We select the time series that most improves model performance as the first predictor. We then repeat the following until we reach a target number of predictors or the model performance levels off: (1) evaluate each *remaining* candidate predictor in combination with predictors already selected for the system and (2) select the candidate that most improves model performance for inclusion in the system. Formally,

$$\boldsymbol{X}_0 \coloneqq \emptyset \text{ and } \boldsymbol{X}_{i+1} \coloneqq \boldsymbol{X}_i \cup \left\{\arg \max_{x \in \boldsymbol{\Omega} \setminus \boldsymbol{X}_i} P(\boldsymbol{X}_i \cup \{x\}, \lambda\,, h; y)\right\} \tag{2.9}$$

where $\boldsymbol{X}_i$ is a set of selected candidate time series at step i; $\boldsymbol{\Omega}$ is a set of all candidate time series; $P(\boldsymbol{X}_i \cup \{x\}, \lambda\,, h; y)$ is the performance metric; y is the gold standard; λ is the smoothing parameter, and h is the threshold for test statistic.

2.3.5 Optimizing early detection of influenza outbreaks in the US

We demonstrate the MEWMA-FFS framework by designing an early detection system for influenza in the US, based on 2010-2016 data. Using national scale ILINet data as the gold standard (described under *Data* below), outbreak events (influenza outbreaks) are defined as ILINet surpassing a specified threshold for at least three weeks.

Candidate predictors are selected to detect the onset of influenza outbreaks as early as possible in a specific number of weeks leading up and following the start of each event.

When selecting candidate predictors, all time series are evaluated using six-fold cross-validation. For each fold, one of the six influenza seasons is held out for testing and the other five are used for training. The candidate model is evaluated by the timing of the alarm relative to the actual start of the event, averaged across the six out-of-sample predictions. To minimize false positives, we set the Average Time between False Signals (ATFS) to 20 weeks and use simulation to find optimization parameter pairs (λ, h) that satisfy this constraint. To reduce the stochasticity of simulation further, we repeat each optimization experiment 40 times, score each predictor by its median rank across the 40 replicates, and select the top scoring predictors for inclusion in the final model.

After selecting optimal combinations of predictors via MEWMA-FFS, we perform two additional rounds of model evaluation. Since the gold standard and predictor data overlap for only six influenza seasons (2010-2016), we used this data twice: first we use six-fold cross validation to select predictors, as described above; second, we use three-fold cross-validation (two seasons held out) to compare the performance of different optimized models. We report the timing of alarms relative to the official start of each event, the proportion of events detected (recall), and the percentage of true alarms over all alarms (precision) across the three folds. In preliminary analysis, we found that the length of training data does significantly impact model performance (Figure 2.5).

Finally, following model construction and comparison on 2010-2016 data, we further evaluate the performance of the best models in comparison to simpler alternative models using true test data from the 2016-2017 influenza season and the fall wave of the 2009 H1N1 influenza pandemic.

2.3.6 Choosing an event threshold and detection window

To speed up the optimization experiments, we tune the event threshold ε and length of detection window T_w. We run optimization experiments using eleven ILINet time series across a range of values for ε and T_w (Figure 2.6). We constrain the T_w so that the start of the window did not precede the lowest observation in the onset of a given outbreak. As in our primary analysis, predictors are selected using 6-fold cross validation and compared via a secondary round of 3-fold cross-validation. We considered ILINet event thresholds ranging from 1% to 2% and detection windows ranging from 4 to 20 weeks surrounding the onset of an event and found that a combination of $\varepsilon = 1.25\%$ and $T_w = 16$ maximizes the timeliness, precision and recall (Figure 2.6).

2.3.7 Assessing the trade-off between run-time and performance

To evaluate the impact of the ATFS on model performance, we run optimization experiments across ATFS values ranging from 5 to 150. In each experiment, predictors are selected and evaluated through cross-validation as described above. For each ATFS value, we run 40 replicates and record their compute time on the Olympus High Performance Compute Cluster [11].

2.3.8 Sensitivity analysis

To evaluate the impact of the training period duration, we run five optimization experiments following the procedures described above, while varying the length of the training time series from 12 years to 4 years: 2004-2016, 2006-2016, 2008-2016, 2010-2016, 2012-2016. To evaluate the importance of including recent data, we run a series of optimization experiments with variable time gaps between the end of a four-year training period and the beginning of a one-year testing period (Figure 2.9).

2.3.9 Alternative models

We compare our optimized early detection systems to three simpler models. All three models were fit via 3-fold cross-validation on 2010-2016 ILINet data, with two seasons held out in each round. When computing performance, we follow the methods described above for the MEWMA-FFS model: We consider only the first alarm in each cluster and assume the same objective function, event threshold, detection window, and ATFS.

Week-based trigger: The model triggers alarms in the same week of every year. Week 34 maximizes the cross-validated performance.

Rise-based trigger: The model triggers alarms as soon as ILINet reports increase for n consecutive weeks. We considered n ranging from 2 to 20 weeks and determined that $n = 4$ maximizes the cross-validated performance.

Univariate-ILINet US: We fit a univariate EWMA model using national level ILINet data as the sole predictor.

2.3.10 Data

The method evaluates candidate data sources based on ability to detect events in a designated *gold standard* data source. Throughout this study, we use CDC national-scale ILINet data as gold standard and consider the following five categories of candidate data: (a) ILINet; (b) NREVSS; (c) Google Trends; (d) Wikipedia access log; (e) athenahealth EHR.

ILINet: The CDC complies information on the weekly number of patient visits to healthcare providers for influenza-like illness through the US Outpatient Influenza-like Illness Surveillance Network (ILINet). Current and historical ILINet data are freely available on FLUVIEW [47]. We use weekly percentage of ILI patient visits to healthcare providers on both national and Health and Human Services (HHS) scales (which are weighted by state population). The national scale time series serve as our gold standard data, and both national and HHS data are considered as candidate data sources during optimization from 07/03/2009 through 02/06/2017.

NREVSS: Approximately 100 public health and over 300 clinical laboratories in the US participate in virologic surveillance for influenza through either US World Health Organization (WHO) Collaborating Laboratories System or the National Respiratory and Enteric Virus Surveillance System (NREVSS). All participating labs issue weekly reports providing the total number of respiratory specimens tested and the percent positive for influenza. These data are publicly available on FLUVIEW [47]. Our optimization considers both national and HHS scale time series of weekly percentage of specimens positive for influenza from 07/03/2009 through 02/06/2017.

GT: Google Correlate [48] and Google Trends [49] are freely-available tools developed by Google that enable users to (1) find search terms correlated with user-provided time series and (2) obtain search frequency time series corresponding to user-provided search terms, respectively. We first applied Google Correlate to national scale ILINet data between 01/04/2004 and 5/16/2009 and retrieved the top 100 matches (Table 2.3). We then applied Google Trends to each of the top 100 search terms to obtain search frequency time series for 07/03/2009 through 02/06/2017. These serve as candidate data sources in our optimization.

Wikipedia: Wikipedia is widely used as a online reference (nearly 506 million visitors per month) [41]. Researchers have demonstrated a correlation between US ILINet and time series of access frequencies for English-language Wikipedia articles relating to influenza [7,41]. Using the Delphi Epidata API [50], we obtained the normalized weekly number of hits for each of 53 influenza-related Wikipedia pages listed in [7] from 07/03/2009 through 02/06/2017 (Table 2.4).

Athena: athenahealth provides cloud-based services for healthcare providers and manages large volumes of electronic health records data. In collaboration with athenahealth, we obtained the following daily data for approximately 71939 healthcare providers across the US from 07/03/2010 to 02/06/2016: the total number of patient visits, the number of influenza vaccine visits, the number of visits billed with a influenza diagnosis code on the claim, the number of ILI visits, the number of visits ordered a influenza test, the number of visits with an influenza test result, the number of visits with a positive influenza test, and the number of visits with a flu-related prescription. We generated 77 time series total for the following seven variables, each aggregated by week and compiled at the national and HHS scale: (1) ILIVisit---the weekly count of ILI visits; (2) ILI%---the ratio of the number of ILI visits and the total number of visits; (3) FluVaccine---the weekly count of visits with an influenza vaccine; (4) FluVisit---the weekly count of visits billed with an influenza diagnosis code on the claim; (5) Positive%---the ratio of the number of visits with a positive influenza test result to the number of visits with a influenza test; (6) FluResult---the number of patient visits with a influenza test result; (7) FluRX---the number of patient visits with a flu-related prescription.

2.4 Results

2.4.1 Early detection from single data sources

We first fit the early detection model to each of the 240 candidate time series individually and assess their ability to anticipate when ILINet will cross a threshold of 1.25%. Performance indicates the average timing of alarms based on six out-of-sample tests, with the range of zero to one corresponding to eight weeks after to eight weeks before the event reaching the threshold 1.25%. The expected performance is highly variable across data sources (Figure 2.1), with ILINet and Google source data generally providing earlier warning than laboratory, EHR and Wikipedia data. The Google Trends time series for 'human temperature' provides the best balance of timeliness, precision and recall (Figures 2.3(A), and 2.7), with an average advanced warning of 14 weeks prior to the CDC's 2% threshold for the onset of the influenza season [51]. National scale ILINet data triggers alarms an average of 11.7 weeks prior to the 2% threshold (Figure 2.3). Several data sources failed to detect any of the seasons, including Wikipedia page views relating to non-seasonal influenza viruses and athenahealth counts of positive influenza tests in HHS regions 8 and 9.

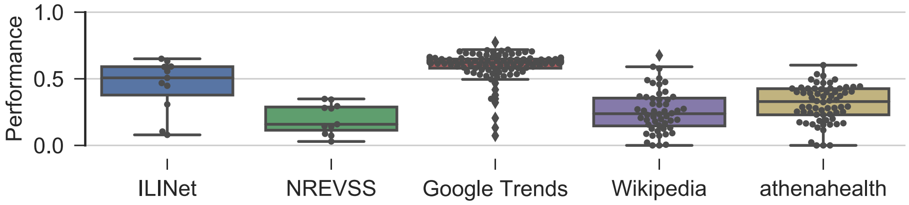

Figure 2.1: Early detection by single data sources, summarized by category. For each of the 240 candidate predictors, we fit a univariate detection model and measured performance by averaging early warning across six-fold cross validation (2010-2016). Emergence events for optimization are defined by an ILINet threshold of 1.25%. The expected performance is highly variable, ranging from 0 to 0.77. A value of one means that the system consistently sounded alarms a full eight weeks prior to the event threshold 1.25%; a value of 0.5 indicates that, on average, the alarms sound at the time reaching the threshold 1.25%; lower values indicate delayed alarms.

2.4.2 Early detection from multiple data sources

We selected optimal combinations of predictors from within each class of data. For CDC ILINet, we considered 11 candidate predictors and found that the optimized system included three time series: ILINet HHS region 7 (Iowa, Kansas, Missouri and Nebraska), ILINet HHS region 5 (Illinois, Indiana, Ohio, Michigan, Minnesota and Wisconsin), and ILINet US (Figure 2.2). Across all replicates, HHS region 7 was selected as the most informative predictor, which alone outperforms the optimized system using multiple NREVSS data sources (Figure 2.2). HHS region 9 and US were not selected in all replicates, and just marginally elevate the performance of HHS region 7. Comparing the optimized internet-source systems (Google Trends and Wikipedia) to optimized EHR (athenahealth) system, we find that the best combination of Google Trends time series---

human temperature, normal body temperature, break a fever, fever cough, flu treatments, thermoscan, ear thermometer---outperforms the others (Figures 2.2, and 2.3(A)).

Across the three-fold out-of-sample tests, the ILINet system detected all six influenza outbreaks with an average advanced warning of 12.7 weeks prior to the CDC's season onset threshold, while the Google Trends system detected 83.3% of outbreaks (five out of six), with an average advanced warning of 16.4 weeks (excluding missing outbreaks) prior to the official threshold (Figures 2.3(A) and 2.7). The other systems each detected four to six of six test seasons (not always the same seasons), with average advanced warning ranging from 9.5 to 14.2 weeks (Figures 2.3(A) and 2.7). Individual ILINet time series generally provide earlier warning than individual EHR and Wikipedia time series. However, performance reverses for optimized multivariate models, with the best ILINet model underperforming both the EHR and Wikipedia models (Figures 2.3(A) and 2.7).

To build multi-category early detection systems, we applied the optimization method to the 'winners' of the previous experiments. That is, we considered the 26 predictors shown on the first five plots of Figure 2.2. The best model includes eight predictors. The top six are all Google Trends: human temperature, normal body temperature, break a fever, fever cough, flu treatments, thermoscan; the remaining two are Wikipedia: orthomyxoviridae and shivering, which only improve the performance of the system marginally (Figure 2.2). None of the ILINet, NREVSS, or EHR time series made the cut. The combined system achieves comparable early warning to the optimized Google Trends system while detecting higher proportion of events with lower number of false alarms (Figure 2.3). Furthermore, it sounds alarms earlier than all three alternative (non-MEWMA) models in four out of six seasons. In 2012-2013 all models provide similar early warning; in 2015-2016, the week-trigger and rise-trigger algorithms signal two and three weeks ahead of our optimized algorithm, respectively (Figure 2.3(B)). The optimized model also produces fewer false alarms than the rise-trigger model and detects a higher proportion of influenza seasons than week-trigger model (Figure 2.3(B)). The MEWMA model using only ILINet data typically lags all other models in signaling events.

When we exclude Google Trends candidates from optimization, the method selects Wikipedia pageviews of flu season as the most informative predictor followed by a combination of EHR, Wikipedia and ILINet time series (Figure 2.2). Expected performance declines slightly without Google Trends data. In three-fold out-of-sample evaluation, the six influenza seasons are detected at an average of 14.8 weeks prior to the CDC's 2% threshold without missing any events (Figure 2.3).

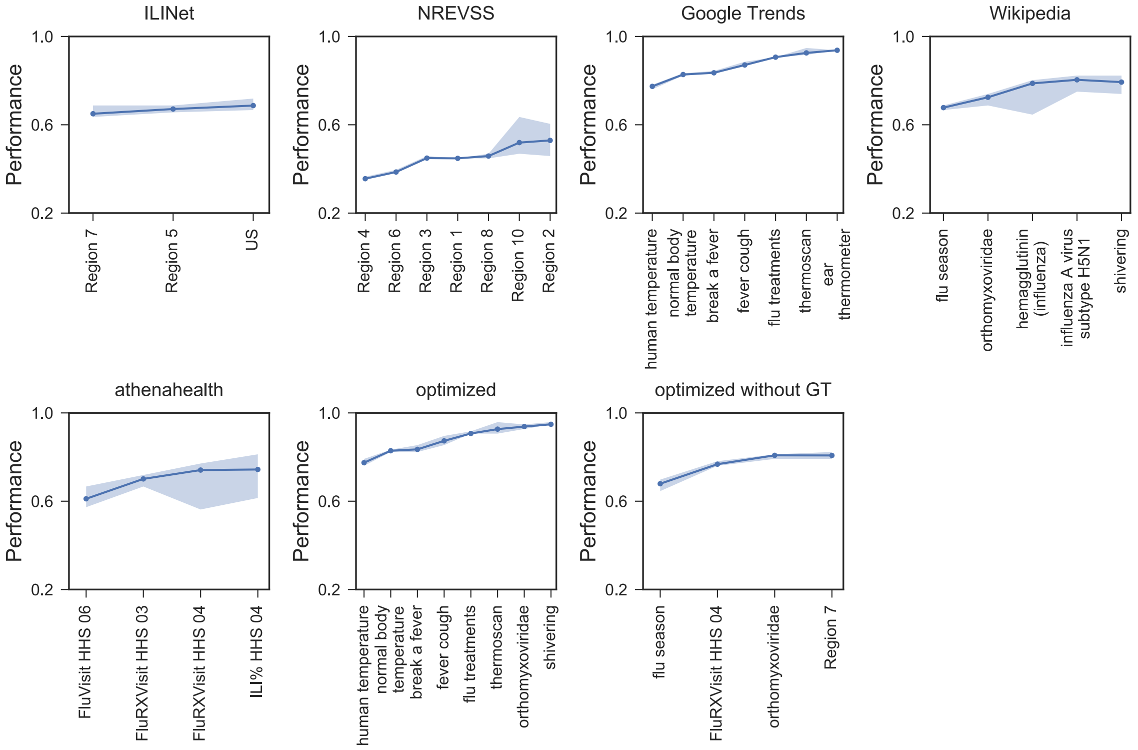

Figure 2.2: Performance curves for early detection systems. Systems were optimized within each data category (ILINet, NREVSS, Google Trends, Wikipedia, and athenahealth) and across all data categories, including and excluding Google Trends. Performance is the average advanced warning within the 16 week detection window surrounding the week when ILINet reaches the event threshold of 1.25%. Performance equal to one indicates that a model consistently signals eight weeks ahead of the event threshold and zero indicating failure to signal within the detection window. Early detection improves as forward selection sequentially adds the most informative remaining data source until reaching a maximum performance. For the optimal system, the first six predictors are Google Trends sources and the remaining two are Wikipedia sources; for the optimal system excluding Google Trends, the top sources are from Wikipedia, athenahealth, Wikipedia and ILINet, in that order.

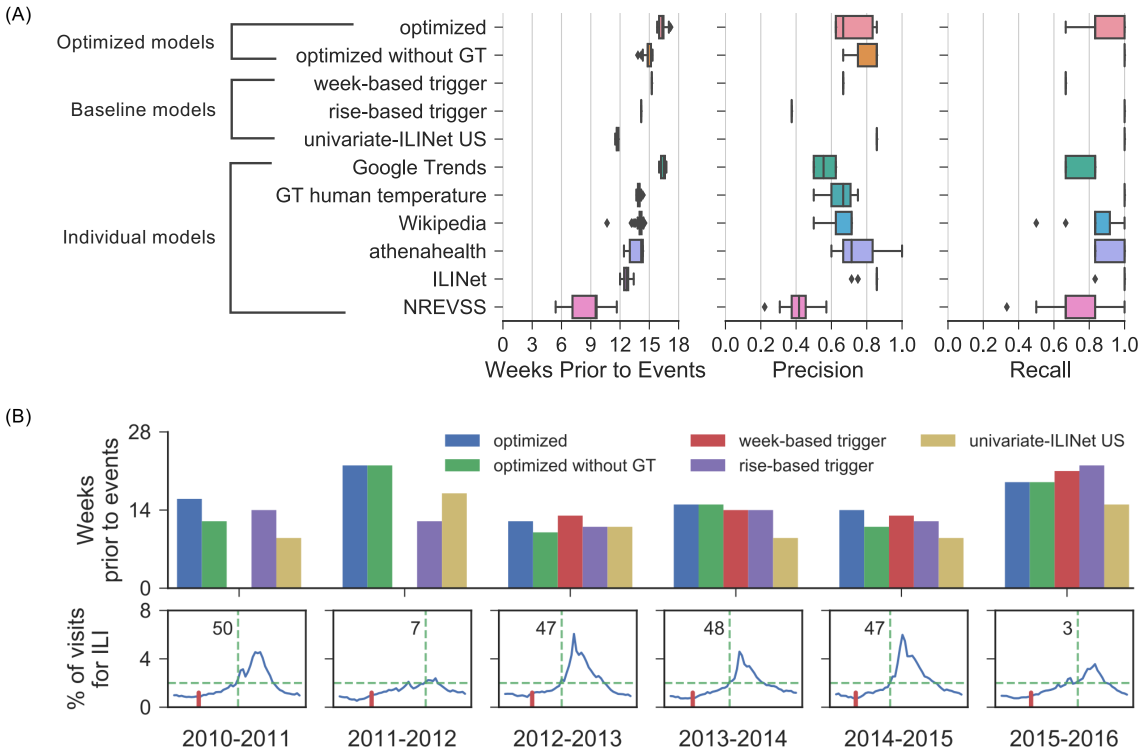

Figure 2.3: Performance of optimized US influenza detection algorithms in three-fold cross validation (2010-2016). (A) Distribution of system performance over six influenza outbreaks across 40 replicates, in terms of the timing of true alarms relative to the official onset of influenza seasons (excluding missed seasons), proportion of alarms indicating actual events (precision), and proportion of events detected (recall). (B) Timing of alarms relative to the official onset of each influenza season. Using US ILINet time series (blue curves) as a historical *gold standard*, the detection models were trained to sound alarms as early as possible in the sixteen weeks surrounding the week when ILINet reaches 1.25%. Bar plot (panel 1) shows the advanced warning provided by out-of-sample alarms in terms of weeks in advance of the CDC's 2% ILINet threshold for declaring the onset the influenza season. Bars not shown indicate missed events. In the lower time series plots, dashed green lines indicate the CDC's seasonal influenza threshold of 2%; numbers indicate the corresponding week of the year; short red lines indicate the timing of the alarms given by the optimized model.

2.4.3 Out-of-sample detection of the 2009 H1N1 pandemic and 2016-2017 influenza season

We further validated our systems using held out ILINet data from two different epidemics. For the 2016-2017 influenza season, the optimized algorithm signaled the start of 2016-2017 season 14 weeks prior to ILINet reaching the CDC's 2% threshold, which outperforms the univariate ILINet model. However, the week-trigger and rise-trigger baselines beat the optimized algorithm by two weeks. For the atypical fall wave of transmission during the 2009 H1N1 pandemic, these two models failed to signal the emerging threat. It emerged much earlier in the year than seasonal influenza (thus tripping up the week-trigger baseline) and at a higher epidemic growth rate (thus outpacing the rise-trigger algorithm) [52]. The optimal system was able to detect the fall wave five weeks prior to ILINet reaching the 2% threshold (Figure 2.4). The univariate ILINet model again lags the best model by several weeks in out-of-sample test. This suggests that our optimized multivariate models are more robust for detecting anomalous influenza threats than the simpler alternatives.

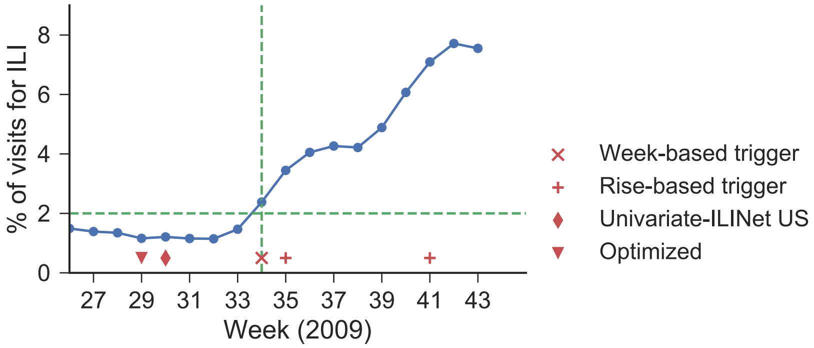

Figure 2.4: Early detection of the 2009 H1N1 pandemic (out-of-sample). The optimized model was trained on 2010-2016 ILINet data, and then tested on US ILINet reports (blue curve) during fall wave of the 2009 H1N1 pandemic. It triggered an alarm (triangle) five weeks prior to ILINet reaching the official epidemic threshold of 2% (dashed lines). Red markers indicate timing of alarms triggered by the optimized and baseline models.

2.4.4 Sensitivity to training period

When we varied the length of the training period from four to twelve years, we selected overlapping sets of optimal predictors, with all five systems including ILINet data for HHS regions 6 and 7 (Table 2.2). The systems detected similar proportions of events. However, the precision (the proportion of true alarms to all alarms) appears to increase with the length of the training period while, surprisingly, the alarms tend to sound later (Figure 2.5). We also found system performance to be fairly insensitive to the gap between the training and testing periods (Figure 2.10), suggesting robust performance with only periodic system updates.

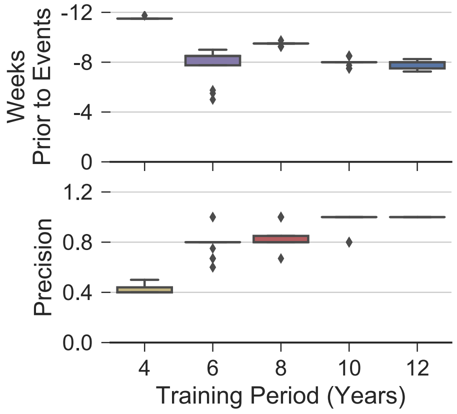

Figure 2.5: Duration of training period impacts early detection. Graphs compare the performance of five systems optimized using continuous training data ranging in length from four to twelve years (each ending in 2016), evaluated via cross-validation on 2012-2016 data. Alarm timeliness (top) unexpectedly declines as the training period increases (maximum likelihood linear regression, P=0.019), while the proportion of true alarms (middle) improves (maximum likelihood linear regression, P=0.000256). Training period does not significantly impact recall (not shown).

2.5 DISCUSSION

This MEWMA-FFS framework is designed to build robust early outbreak detection systems that harness a variety of traditional and next generation data sources. For influenza outbreaks in the US, we identified a combination of freely available internet-source data that robustly detects the start of the season an average of 16.4 (SD 3.3) weeks in advance of the national surveillance threshold (ILINet reaching 2%). This is five weeks earlier than previously published early detection algorithms based on ILINet and Google Flu Trends data [20,21]. In a retrospective out-of-sample attempt to detect the fall wave of the 2009 H1N1 influenza pandemic, the optimized multivariate algorithm provided the earliest warning among the competing models. However, it sounded an alarm only five weeks prior to ILINet reaching the national 2% threshold. The shorter lead time may stem from the anomalously rapid growth of the 2009 pandemic. Across the six influenza seasons between 2010 and 2017, ILINet took an average of 9.4 weeks to increase from 1.25% to 2%, with a minimum of six weeks in seasons 2012-2013 and 2014-2015; in the fall of 2009, this transpired in a single week (week 34).

Public health surveillance data (e.g., ILINet and NREVSS) can detect emerging influenza seasons on their own, but a combination of eight Google query and Wikipedia pageview time series provided earlier warning across all eight epidemics tested. Although we cannot definitively explain the performance of internet data, we note that 59% of flu-related Wikipedia English pageviews come from countries outside the US, including the United Kingdom, Canada, and India [41]. Perhaps earlier influenza seasons elsewhere provide advanced warning of imminent transmission in the US. The utility of Google and Wikipedia data may also stem from their large and diverse user bases and their immediate use following symptoms relative to seeking clinical care [53]. NRVESS is among the mostly costly and time lagged data sources; it performs poorest when considered individually and is never selected for inclusion in combined early detection systems. However, NRVESS provides critical spatiotemporal data for detecting and tracking novel viruses, including pandemic and antiviral resistant influenza, and informing annual vaccine strain selections. Thus, we speculate that NRVESS might rank among the most important sources when designing systems for virus-specific influenza nowcasting and forecasting objectives.

We emphasize that these models are not designed to forecast epidemics, but rather to detect unexpected increases in disease-related activity that may signal an emerging outbreak [17]. Early warning provides public health agencies valuable lead time for investigating and responding to a new threat. For seasonal and pandemic influenza, such models can expedite targeted public health messaging, surge preparations, school closures, vaccine development, and antiviral campaigns. Influenza forecasting models potentially provide more information about impending epidemics, including the week of onset, the duration of the season, the overall burden, and the timing and magnitude of the epidemic peak [54–56]. However, they are typically not optimized for early warning or for detecting outbreaks that are anomalous in either the timing or pace of expansion.

Our conclusions may not be readily applied to influenza detection outside the US or to other infectious diseases. However, the general framework could be similarly deployed to address such challenges. Even for influenza outbreaks in the US, our results pertain to only early detection of influenza outbreak activity as estimated from ILINet, and stem from only six seasons of historical data. If we changed the optimization target (i.e., gold standard data) to an EHR or regional ILINet source, the resulting data systems and corresponding performances may differ considerably. Furthermore, as alternative data and longer time series become available, the optimal systems could potentially improve. Early detection systems should therefore be regularly reevaluated and tailored to the specific objectives and geopolitical jurisdictions of public health stakeholders, and our optimization framework can facilitate easy and comprehensive updates.

This approach requires domain-knowledge in the selection of candidate data sources. Next generation proxy data should be relevant to the focal disease and population, such as symptom or drug related search data. Climate and environmental factors may prove predictive for directly transmitted and vector borne diseases, and may be a promising direction for enhancing the early detection systems developed here. This black box approach can select data sources with spurious or misleading relationships to the gold standard data. Thus, it may be prudent to screen data sources before and after optimization that are unlikely to correlate reliably with the target of early detection.

We implemented this MEWMA-FFS framework as a user-friendly app in the Biosurveillance Ecosystem (BSVE) built by the US Defense Threat Reduction Agency (DTRA) [57]. Military bioanalysts can now use it to evaluate and integrate diverse data sources into targeted early detection systems for a wide range of infectious diseases worldwide. The versatility of this plug-and-play method stems from two assumptions: (1) it simply scans for deviations from underlying distributions rather than modeling a complex epidemiological process, and (2) it does not require seasonality, just historical precedents with which to train the model. We can now more easily harness the growing volumes of health-related data to improve the timeliness and accuracy of outbreak surveillance and thereby improve global health.

2.6 Supplemental information

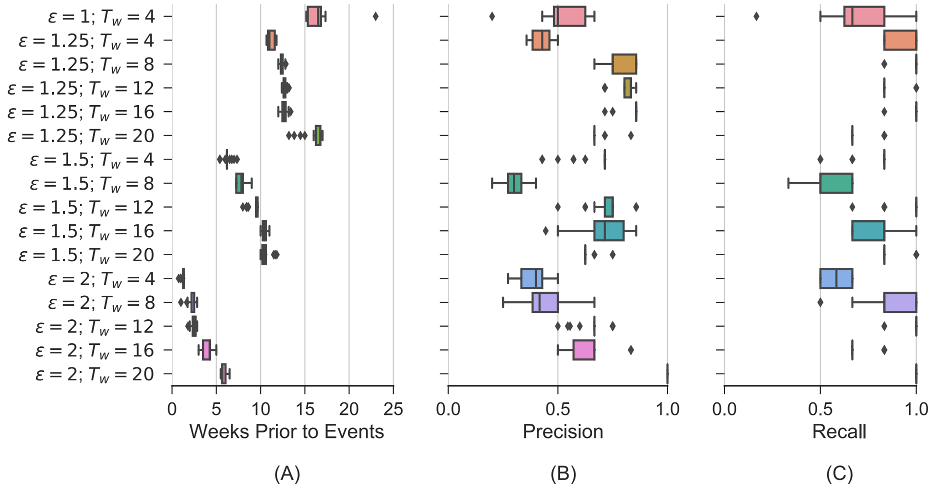

Figure 2.6: Comparison of system performances with different pairs of event threshold ε and detection window T_w in three-fold cross validation (2010-2016). Distribution of average system performance over six influenza seasons across 40 replicates, in terms of the timing of true alarms(excluding missed seasons), proportion of alarms indicating actual events (precision), and proportion of events detected (recall).

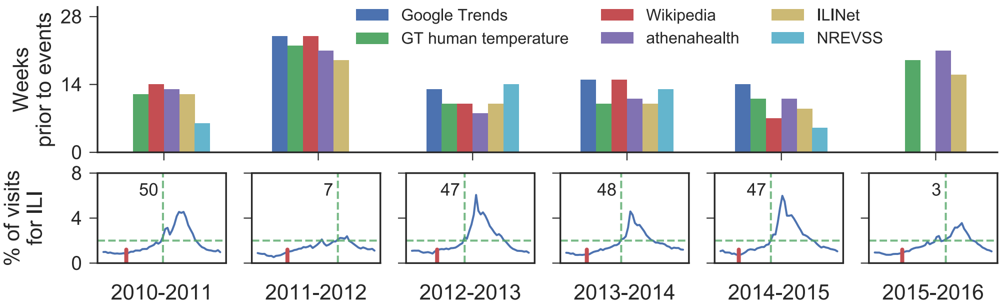

Figure 2.7: Out-of-sample detection of US influenza seasons by single source and single category early warning systems. Using US ILINet time series (blue curves) as a historical gold standard, the detection models were optimized to sound alarms as early as possible in the sixteen weeks surrounding the threshold 1.25% for optimization. The bar plot (panel 1) shows the alarm timing for each influenza season from 2010-2016 relative to the official ILINet threshold of 2%. Bars not shown indicate missed events in early detection, while positive values show alarms are triggered prior to the official start of each influenza season. In panel 2, horizontal green dashed lines represent the threshold of 2%, while vertical green dashed lines indicate the onset of influenza seasons according to the threshold of 2%; numbers indicate the corresponding week of the year; red short lines show alarm timings for flu seasons from the optimized model.

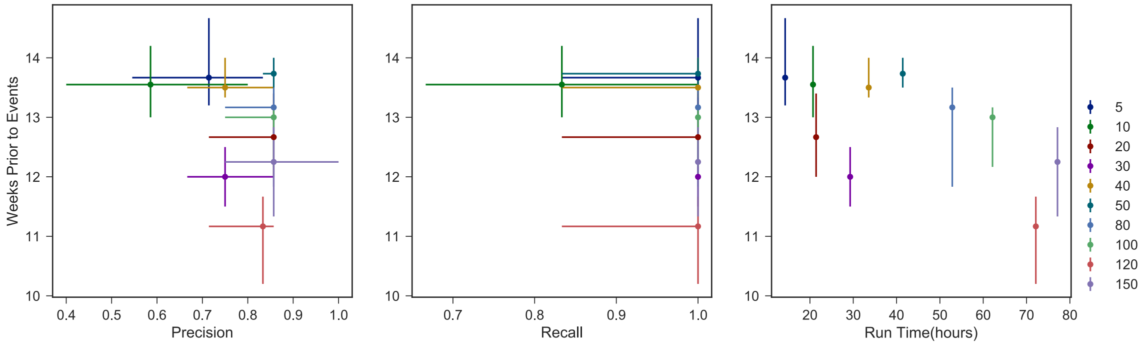

Figure 2.8: The trade-off between timeliness, and precision, recall, running time. Each system was optimized using different values of ATFS. The three plots show the trade-off between alarms timings and the proportion of alarms indicating actual events (precision), proportion of events detected (recall), and running time of each optimization with 40 repeats running in parallel, respectively. Each run selected different combinations of predictors (Table 2.1) and detected influenza emergence an average of 11-14 weeks prior to the official onset of influenza seasons. There is a weak trade-off between timeliness and precision and minimal trade-off between timeliness and recall. The precision is always below 0.9 while recall is equal to one for most of values of ATFS. This is because we consider the timing of only the first alarm in a cluster; the ATFS is expected to impact the total number of alarms but not necessarily the number of alarm clusters [17]. Meanwhile, a larger value of ATFS requires longer running time for optimization. An optimization experiment with ATFS set to 50 (the value that maximizes timeliness and precision) requires twice the run time of an experiment using ATFS 20; however, the gain is only one additional week of early warning. Thus, it is valuable to balance performance and compute time when setting ATFS for optimization.

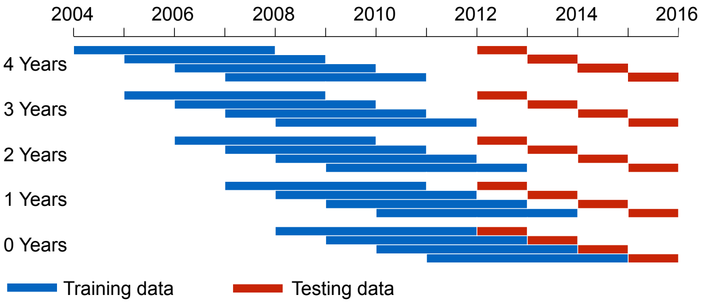

Figure 2.9: Diagram of training and testing periods used in sensitivity analysis.

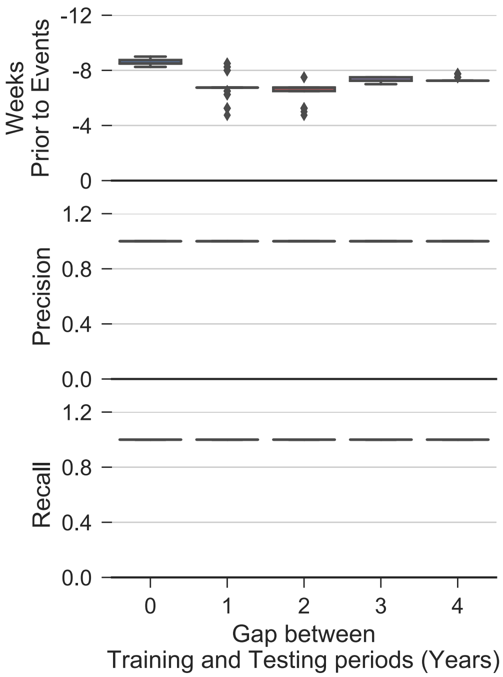

Figure 2.10: Sensitivity to the training period. Each of five systems was optimized using training and testing periods diagrammed in Figure 2.9. The three graphs show performance in terms alarm timing (top), proportion of alarms that correspond to actual events (middle), and proportion of events detected (bottom). Gap between testing and training periods does not appear to significantly impact performance.

Table 2.1: Time series selected for early detection systems across different values of ATFS. Time series are listed in order of selection, assuming an ILINet threshold of 1.25% for optimization.

Value of ATFS									
5	10	20	30	40	50	80	100	120	150
HHS 7	HHS 7	HHS 7	HHS 7	HHS 7	US	HHS 7	US	US	US
	HHS 6	HHS 5	US	US	HHS 4	HHS 6	HHS 4	HHS 9	HHS 4
	HHS 2	US	HHS 6	HHS 10	HHS 10	HHS 4	HHS 6	HHS 6	HHS 6
				HHS 4	HHS 6	US	HHS 7	HHS 5	HHS 7
				HHS 6				HHS 1	
				HHS 8					

Table 2.2: Data sources selected for early detection systems across variable length training periods. Time series are listed in order of selection, assuming an ILINet event threshold of 1.25%.

Model training period				
2004-2006	2006-2016	2008-2016	2010-2016	2012-2016
HHS 5	US	HHS 7	HHS 7	HHS 3
HHS 7	HHS 6	HHS 1	HHS 9	HHS 7
HHS 9	HHS 7	HHS 6	HHS 6	HHS 10
HHS 6	HHS 9			HHS 2
HHS 8	HHS 8			HHS 6
	HHS 2			US

Table 2.3: Candidate Google Trends data sources for early detection of seasonal influenza. Optimization experiments evaluated 100 time series based on each of these search terms.

Google search terms			
pneumonia	flu headache	signs of flu	how long does flu last
treating flu	low body	early flu symptoms	normal body temperature
flu report	get over the flu	influenza type a	how long is the flu contagious
flu duration	how long flu	symptoms of flu	incubation period for flu
after the flu	flu how long	get rid of the flu	how to treat the flu
flu cough	flu coughing	break a fever	flu contagious period
flu fever	having the flu	type a influenza	ear thermometer
treat the flu	i have the flu	treatment for flu	how to get rid of the flu
flu last	flu contagious	human temperature	influenza incubation period
flu vs. cold	dangerous fever	when you have the flu	symptoms of bronchitis
the flu	cold versus flu	signs of the flu	what to do if you have the flu
cold and flu	flu in children	taking temperature	over the counter flu
flu type	remedies for flu	if you have the flu	over the counter flu medicine
flu germs	contagious flu	do i have the flu	how long is the flu
flu recovery	exposed to flu	symptoms of the flu	incubation period for the flu
cold vs. flu	is flu contagious	treating the flu	how long does the flu last
thermoscan	have the flu	flu and fever	how long does the flu
flu or cold	oscillococcinum	flu and cold	how long contagious
flu lasts	flu medicine	fight the flu	how long is flu contagious
flu care	flu treatments	reduce a fever	how to reduce a fever
flu length	flu complications	upper respiratory	fever dangerous
treat flu	influenza symptoms	high fever	flu treatment
cure flu	cold vs flu	flu children	medicine for flu
cure the flu	braun thermoscan	the flu virus	cold symptoms
flu vs cold	fever cough	how to treat flu	is the flu contagious

Table 2.4: Candidate Wikipedia data sources for early detection of seasonal influenza. Optimization experiments evaluated 53 time series based on access frequency for each of these Wikipedia articles.

Wikipedia articles			
Influenza A virus subtype H5N1	Nausea	Antiviral drugs	Influenza-like illness
Influenza A virus subtype H7N2	Headache	Rhinorrhea	Influenzavirus A
Influenza A virus subtype H7N3	Malaise	Rimantadine	Canine influenza
Influenza A virus subtype H7N7	Chills	Equine influenza	Swine influenza
Influenza A virus subtype H9N2	Influenza	Paracetamol	Influenzavirus C
Influenza A virus subtype H7N9	Myalgia	Common cold	Orthomyxoviridae
Influenza A virus subtype H1N1	Cough	Nasal congestion	Influenza vaccine
Influenza A virus subtype H10N7	Fever	Sore throat	Viral pneumonia
Influenza A virus subtype H1N2	Vomiting	Fatigue (medical)	Influenza B virus
Influenza A virus subtype H2N2	Shivering	Gastroenteritis	Viral neuraminidase
Influenza A virus subtype H3N8	Flu season	Avian influenza	Influenza pandemic
Influenza A virus subtype H3N2	Oseltamivir	Cat flu	Influenza prevention
Hemagglutinin (influenza)	Zanamivir	Influenza A virus	Human flu
Neuraminidase inhibitor			

Chapter 3: Optimizing data sources for early detection of large dengue outbreaks

3.1 ABSTRACT

Dengue has been endemic in populations throughout Latin American, the Pacific islands, and Southeast Asia over decades, causing large outbreaks during some but not all years. Early and accurate detection of emerging epidemics can be critical to effective public health intervention. Here, we extend a multivariate early detection framework for detection of emerging waves of dengue transmission as early as possible that leverages a combination of historic incidence, climate, and search query data. We apply signal processing methods to detect anomalous dengue-related activity and feature selection algorithms to evaluate hundreds of candidate data sources for inclusion in the detection model. Optimal models for each study area contain fewer than eight of more than one hundreds possible candidate predictor time series, and outperform baseline models including only dengue incidence data from public health surveillance systems. We found that the framework is very sensitive to locations, and an optimal set of predictor time series derived from one location may not be applied to other locations directly. This study not only contributes a framework to select small subsets of data sources from among hundreds of candidates to improve the early detection of large emerging and re-emerging epidemics, but also proves flexibility of the framework on different locations.

3.2 INTRODUCTION

Dengue, a mosquito-borne infectious disease, has emerged as a public health problem since the 1960s, and remains endemic in many tropical and subtropical countries, with about four billion people at the risk of dengue infection [58]. Between 1990 and 2013, the number of symptomatic dengue infections doubled every 10 years, and approximate 60 million people are infected symptomatically by dengue virus per year on average with nearly 10,000 deaths from the infection [59]. Not only dengue incidence keeps increasing over years, but explosive outbreaks are occurring in some years. For example, Puerto Rico has experienced four large outbreaks since 1990 with about 20,000 suspected cases on average, while in other years, the number of suspected cases were 3,000 to 9,000 [60]. In Mexico and Peru, the number of dengue cases was various significantly between years as well ranging from several thousand to over 200,000 infections [61], and from 6,000 to more than 60,000 cases [62], respectively. The variation and non-seasonality of dengue outbreaks make the detection of large dengue outbreaks challenging.

Public health agencies need reliable surveillance systems to prevent or slow down the spread of dengue outbreaks in a timely manner. For instance, with sufficient early warnings, public health officials are able to be well-prepared -- determining when and where to distribute vaccines and antivirals, and implementing effective interventions, such as spraying insecticides. Traditional passive surveillance systems for dengue usually depend on voluntary reports from healthcare facilities [63,64], and an outbreak is announced if the number of dengue cases surpasses the 75th percentile of the distribution of weekly number of dengue cases based on historical data [60]. However, many past outbreaks have shown that there was always a delay of days to weeks between symptoms onset and case reporting. It also takes extra days for public health agencies to collect and publish the reports. Moreover, passive surveillance systems tend to underreport the total number of cases since many dengue infections have no or mild symptoms, and thus do not seek medical treatment [65–68]. Consequently, early warnings received from traditional systems are not actually early, and usually in the exponential growth phase of an outbreak. Therefore, it is urgent to develop effective active surveillance systems which are able to detect large dengue outbreaks early and accurately.

Lots of statistical models have been developed to forecast dengue outbreaks based on either historical incidence data or climate data. Some of them focused on short-term forecast (<= 3 months) and predicted dengue incidence directly in monthly or weekly level [69–75]. Early warning systems were also built to identify large dengue outbreak years starting from March based on only incidence and climate data in the preceding months from October to December, which were not able to detect if the outbreak has been started or not [76,77]. In addition, Bowman *et al*. applied logistic regression to identify large dengue outbreaks from small ones in real-time using only climate data and epidemiological data without feature optimization [78]. During the dengue forecasting competition hosted by several departments in the US Federal Government in 2015 [79], some teams developed non-statistical or statistical models to forecast the peak height and week based on historical incidence data which outperformed baseline models [80–82]. To predict peaks of outbreaks, the forecasting usually initializes after the start of an outbreak, and thus is not able to predict the start of an outbreak in advance.

Considering in the big data era, there are plenty of data sources available from not only public health surveillance systems and climate stations, but also the Internet, such as Google, Twitter, Wikipedia. Previous research has shown that Internet data is informative for infectious diseases surveillance. For example, Google search query data has been used to monitor or forecast infectious disease dynamics in real time, such as dengue [44,83–88], and influenza [40,89–104]; influenza forecasting and situational awareness using Wikipedia data also achieved good performance [7,41,95,96,105]; HealthMap integrates various freely available electronic media sources to obtain the status of infectious diseases globally [106–109]. However, Internet data relevance to infectious diseases are sensitive to human behaviors, such as panic-induced searching caused by disease-relevance media news [43,44]. Therefore, to include Internet-sourced data in disease surveillance models, it is essential to optimize and evaluate those data throughout multiple years to ensure a robust performance.

In a recent study, we developed a hierarchical early detection framework that can be applied to detect not only re-emerging but also emerging outbreaks in real time [110]. The framework includes an anomaly detection model and a forward feature selection (FFS) algorithm. The anomaly detection model is a Multivariate Exponentially Weighted Moving Average (MEWMA) method, which detects anomalies in a target time series by monitoring multiple predictor time series. The model assumes that observations within non-outbreak periods follow a multivariate normal distribution (null distribution). The model sounds an alarm if an observation is away from the multivariate normal distribution (See Methods for details). The feature selection algorithm determines which time series should be included as predictors in the MEWMA model. False alarms are constrained by a predefined parameter Average Time between False signals (ATFS). The framework performed better than baseline models in detecting the second wave of the 2009 flu pandemic, with alarms being triggered five weeks in advance of the official start of flu seasons [110].

Parameter optimization is a problem of choosing a set of optimal parameters for a statistical or machine learning model. There are five types of parameter optimization methods in general: grid search, random search, gradient-based optimization, evolutionary optimization and Bayesian-based optimization. Grid search is an exhaustive searching through predefined parameter spaces and evaluate all combinations of parameters values via an objective function [111], and random search is to select combinations of parameter values randomly, instead of exhaustively, for evaluation [112]. However, both searching methods are time-consuming because of the large size of potential combinations. Consequently, other optimization methods are developed. Gradient-based optimization is the most popular method for parameter optimization in machine learning, which is a first-order iterative optimization algorithm [113]. To apply this method, an objective function must be differentiable. Evolutionary optimization is a group of methods based on the biological concept of evolution, such as mutation, recombination, reproduction and selection [114,115]. Each combination of parameter values is considered an individual in a population of combinations, and the quality of each individual is defined by an objective function. Evolution of parameters occurs among individual combinations with high quality. Bayesian-based optimization is a set of algorithms for noisy black-box objective functions [116]. The optimization starts with predefined prior distributions for each parameter, and the prior distributions are updated to form posterior distributions over an objective function. The posterior distributions are used to sample new combinations of parameter values, and get updated over the objective function until no improvement is observed in model performance. Both evolutionary and Bayesian-based algorithms do not require derivatives of an objective function.

Here, we modified the MEWMA-FFS framework to detect large dengue outbreaks: (1) To make the anomaly detection model have more power to distinguish large outbreaks from small ones, we introduce a parameter *baseline threshold* to the model, and only observations between *baseline threshold* and *event threshold* are counted in the null distribution; (2) To decrease computational time complexity, we use a penalty term in the objective function, instead of the predefined ATFS, to constrain false positive alarms in order to get rid of the simulation-based parameter optimization. Instead, the Tree-structured Parzen Estimators (TPE) [117], a Bayesian-based optimization algorithm, is applied to estimate model parameters, including *baseline threshold*, *smoothing parameter* and *statistical threshold*. We applied the upgraded framework to detect large dengue outbreaks in Mexico, San Juan metropolitan area in Puerto Rico, and Iquitos metropolitan area in Peru from 2004-2017, 2004-2013, and 2004-2013, respectively, with candidate time series from Google Trends, climate stations and public health surveillance systems. Our study shows that the early detection framework is very sensitive to locations, and different optimal combinations of predictor time series are selected for each study area. With limited data in out-of-sample evaluation, optimized early detection systems outperform three baseline models which only use an incidence time series from public health surveillance systems. We also show that combinations of time series optimized for one specific location is not informative for other locations.

3.3 MATERIALS AND METHODS

3.3.1 Early detection framework

The early detection framework includes two layers: (i) anomaly detection layer that is responsible to detect outbreaks, (ii) data optimization layer to select which candidate time series should be included in the anomaly detection layer as predictors.

3.3.1.1 Anomaly detection layer

The model implemented in anomaly detection layer is a Multivariate Exponentially Weighted Moving Average (MEWMA) method. In the model, we define one data source as target time series y, and all others as predictor time series $\boldsymbol{X}$, where $\boldsymbol{X} = [X^1, X^2, \dots X^M]^T$, and M is the number of predictor time series.

Event threshold ϵ is a predefined value within the range of y that is used to label time periods belonging to anomalies. All observations above ϵ on y are considered as outbreaks, and the corresponding time periods are outbreak periods. To make the model identify large outbreaks from small ones, we introduce a parameter *baseline threshold* ϵ_0 to the model, and only time periods $\boldsymbol{T}$ with corresponding observations on y falling between ϵ and ϵ_0 are defined as non-outbreak periods. That is, $\boldsymbol{T} = \{t: \epsilon_0 < y_t \leq \epsilon\}$. We project $\boldsymbol{T}$ to predictor time series $\boldsymbol{X}$ to obtain observations within non-outbreak periods. The model assumes all observations within non-outbreak periods follow a multivariate normal distribution:

$$\boldsymbol{X_T} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \tag{3.10}$$

where $\boldsymbol{\mu}_0$ is the mean vector and $\boldsymbol{\Sigma}_0$ is the covariance matrix.

Next we compute the exponentially weighted moving average $\boldsymbol{S}_t$ corresponding to each time point on predictor time series $\boldsymbol{X}$ using Equation (3.11) [45].

$$\boldsymbol{S}_t = \begin{cases} \boldsymbol{X}_t, & \text{if } t = 1 \\ \max[\boldsymbol{0}, \lambda(\boldsymbol{X}_t - \boldsymbol{\mu}_0) + (1-\lambda)\boldsymbol{S}_{t-1}], & \text{if } t > 1 \end{cases} \tag{3.11}$$

where λ is the smoothing parameter determining the degree of weighting decrease ($0 < \lambda < 1$). With a higher λ, the data prior to t decays faster. $\boldsymbol{X}_t$ is a column data vector at time point t in predictor time series $\boldsymbol{X}$.

The covariance matrix of $\boldsymbol{S}_t$ equals to

$$\boldsymbol{\Sigma}_{\boldsymbol{S}_t} = \left(\frac{\lambda[1-(1-\lambda)^{2t}]}{2-\lambda}\right)\boldsymbol{\Sigma}_0 \tag{3.12}$$

When computing the MEWMA statistic, we use the asymptotic covariance matrix as $t \to \infty$ [118], that is,

$$\boldsymbol{\Sigma}_{\boldsymbol{S}_\infty} = \left(\frac{\lambda}{2-\lambda}\right)\boldsymbol{\Sigma}_0 \tag{3.13}$$

Based on Hotelling's multivariate control-chart procedure [119], MEWMA model sounds an alarm as soon as

$$E_t = \boldsymbol{S}_t^T \boldsymbol{\Sigma}_{\boldsymbol{S}_\infty}^{-1} \boldsymbol{S}_t > h \tag{3.14}$$

where h is a specified statistic threshold. In this model, only three parameters need to be estimated, including baseline threshold ϵ_0, smoothing parameter λ, and statistic threshold h, no matter how many predictor time series are included.

3.3.1.2 Data optimization layer

To evaluate and select which predictor time series should be included in the MEWMA model, we implement Forward Feature Selection (FFS) algorithm in the framework. The procedure starts with no predictor time series in the model, and evaluates the one-component subsets $\{X^1\}, \{X^2\}, \dots, \{X^M\}$. Then the best individual predictor time series $\{X^m\}$ will be included in the model. Next, the algorithm evaluate each two-component subset consisting of $\{X^m\}$ and one other time series from the remaining

$M-1$ input predictors. And the model is updated to include the two-component subset with the best performance. The procedure is repeated until the model reaches a predefined number of predictor time series (Table 3.1).

Table 3.1: Forward feature selection procedure

```
Algorithm
Require: A set of all candidate time series $\mathbf{\Omega}$
Require: Pre-define a size of predictor time series in the model $K$
    Create an empty set: $X^{(k)} = \{\emptyset\}, k = 0$
    while $k < K$ do
        $X^{m} = \arg\max_{X \in \mathbf{\Omega} \setminus \boldsymbol{X}^{(k)}} J(\boldsymbol{X}^{(k)} \cup \{X\})$
        Update $\boldsymbol{X}^{(k)} \leftarrow \boldsymbol{X}^{(k)} \cup \{X^{m}\}$
        $k += 1$
    end while
```

3.3.2 Objective function

To measure the performance of each predictor time series in feature selection, we define a detection window prior to each outbreak to quantify how early an alarm is triggered, and alarms outside detection windows are considered as false alarms. The performance of each predictor is computed as:

$$J(\boldsymbol{X}, \epsilon_0, \lambda, h; y) = \frac{1}{N} \sum_{n=1}^{N} \frac{\Delta W_n}{W_n} \left(\frac{1}{\theta + 1} \right)^{\omega} \tag{3.15}$$

Here, N is the total number of events labeled in target time series y; W_n is the length of detection window for outbreak n which is prior to the start of an outbreak, and ΔW_n is the time difference between the first true alarm for outbreak n and the onset of the corresponding outbreak. We define $\Delta W_n = 0$ if no true alarm is triggered for outbreak n. In addition, if there is no outbreak during a time period, we make $\Delta W_n = W_n$ and $N = 1$. To constrain false alarms, we add a penalty parameter $\omega \in [0, \infty)$ to the total number of false alarms θ. A penalty of 0 indicates no penalty to false alarms. The value of the objective function ranges from 0 to 1, with a score of 1 indicating alarms are always triggered at the start of detection windows without false alarms.

3.3.3 Parameter estimation

Given that the objective function is not differentiable, we apply the Tree-structured Parzen Estimators (TPE) algorithm to estimate parameters of the framework, including baseline threshold ϵ_0, smoothing parameter λ, and statistic threshold h. TPE is a Bayesian-based optimization algorithm, which does not require to specify initial guesses for parameters. We use the TPE algorithm implemented in a Python library 'hyperopt' [120] to run the optimization. We define a uniform distribution as the initial distribution for each parameter:

$$\epsilon_0 \sim \mathcal{U}(0, \epsilon); \quad \lambda \sim \mathcal{U}(0, 1); \quad h \sim \mathcal{U}(0, 800) \tag{3.16}$$

During the first 20 iterations, we apply random search to get initial combinations of parameter values. Each combination is evaluated according to the objective function. All combinations are divided into two groups using the default fraction in the package: the first group contains the ones with higher objective values and the second group includes all others. The density function of each group is estimated using a nonparametric method -- parzen-window density estimation. The next step is to sample new combinations of parameter values that are more likely to be in the first group and less likely to be in the second group. The new combination with the highest improvement is then used to update the density function. The process is repeated until achieving 1000 iterations.

3.3.4 Building early detection systems

We apply the MEWMA-FFS framework to build early detection systems for three study areas: San Juan metropolitan area in Puerto Rico, Iquitos metropolitan area in Peru, and the country of Mexico. Data from each area is split into training and testing periods (Table 3.7).

We use dengue weekly incidence data collected from public health surveillance systems as *target* time series to label large dengue outbreaks, and determine the length of their detection windows. To remove noises and reveal the trend of dengue infection, we smooth each time series using Exponentially Weighted Moving Average (Equation (3.17)) prior to labeling large dengue outbreaks (Figure 3.3). A smoothing parameter of 0.2 is applied to the target time series for San Juan and Iquitos, and a value of 0.4 is applied to that of Mexico. A large outbreak is defined as weekly incidence surpassing the 75th percentile of the distribution of historical weekly number of dengue cases based on training data (event threshold ϵ) in four consecutive weeks, and the peak height of the outbreak is 1.4 times of the threshold (Figure 3.3). The start of a detection window is defined as the first week when weekly incidence increases for three consecutive weeks during the onset of a given outbreak.

$$A_t = \lambda C + (1 - \lambda)A_{t-1} \tag{3.17}$$

To select optimal combinations of predictor time series, we evaluate all time series using cross-validation on training data. Each fold includes one year of data from the spring of current year to that of next year. In each round, one fold is held for testing and all others are held for model training, and the performance of each predictor time series is averaged across all testing folds in one single experiment. In addition, to ensure the robustness of early detection systems, we repeat each experiment 20 times and the final performance of each predictor time series is determined by the median, and the one with the best median performance is included in the system.

To choose the best penalty parameter ω to constrain false alarms, we vary the value of ω from 0 to 5 while optimizing predictor time series (Figure 3.4). The best value of ω is determined based on three metrics – alarm timing (the ratio of the distance between the first true alarm and the onset of an outbreak to the length of the detection window for the corresponding outbreak), precision (the proportion of true alarms over all alarms) and recall (the proportion of large outbreaks detected). A timing of 1 indicates that an alarm is triggered at the start of a detection window, and 0.5 shows that a model sound an alarm at the middle of a detection window. Next, we validate optimal combinations of predictor time series and penalty parameter ω using data from testing period, and the performance of each system is compared according to the three metrics described above across 20 replicates, in which only alarms triggered in over 50% of all 20 replicates are kept in calculating model performance.

3.3.5 Baselines

We compare early detection systems constructed for each study area with three simple baselines: two pure data-driven baselines (week-based trigger and rise-based trigger) and one model-driven baselines (univariate version of the MEWMA model). Each baseline only use target data (weekly dengue incidence data from public health surveillance systems), and is trained and tested using the same time periods as the early detection systems described in Table 3.7. We consider only the first alarm in each cluster. The same objective function, event threshold, detection window as described above are used in three baselines. *Week-based trigger* sounds alarms the same week w each year, and the week number w maximizing alarm timing and the number of true alarms in training data is selected for each region as the week-based trigger (Table 3.2). *Rise-based trigger* triggers alarms when weekly dengue incidence keeps increasing in n consecutive weeks. We considered n ranging from 2 to 20 weeks and the value of n that maximizes the alarm timing and the number of true alarms in training period is chosen as the rise-based trigger (Table 3.2). *Univariate version of the MEWMA model* only uses weekly dengue incidence from surveillance systems as a predictor to detect large dengue waves.

Table 3.2: Parameter values in each baseline for corresponding locations

Baseline	Mexico	San Juan	Iquitos
week-based trigger	$w = 16$	$w = 25$	$w = 39$
rise-based trigger	$n = 6$	$n = 7$	$n = 4$

3.3.6 Data

We collect time series data from the following four sources, and convert each time series into three candidate predictors: (i) the level (value of the time series in the trailing week); (ii) the slope (first difference over trailing two weeks); (iii) the acceleration (second difference over trailing three weeks).

(1) Reported weekly dengue incidence from passive surveillance systems (Table 3.8). We collect number of reported dengue cases for San Juan metropolitan area, Puerto Rico and Iquitos metropolitan area, Peru from the website of Epidemic Prediction Initiative [79], and for the country of Mexico from the website of Ministry of Health of Mexico [121].

(2) Daily climate data from the website of National Oceanic and Atmospheric Administration (NOAA) [122]. NOAA integrates climate data from land-based climate stations across the world that have been subjected to a common suite of quality assurance reviews. It includes numerous climate factors, such as temperature, rainfall etc. We determine to choose three climate factors as candidate predictors, including temperature [123,124], precipitation [125,126], and relative humidity [127–129], which can affect dengue infection by influencing biology of mosquitoes. We retrieve daily max

temperature, min temperature, average temperature and precipitation from GHCN (Global Historical Climatology Network) database [130], and daily relative humidity from ISD (Integrated Surface Database) database [131] across climate stations around corresponding regions. We aggregate daily data into weekly scale by taking medium of each week (Table 3.8).

(3) Weekly sea surface temperature data from the website of National Centers for Environmental Prediction (NCEP) [132]. Many studies have shown that El Niño and its effect on local meteorological conditions influence inter-annual variability in dengue transmission [133–141], and sea surface temperature is a key indicator of El Niño. Therefore, we include weekly sea surface temperature (SST) and SST anomalies (SSTA) from four Nino regions on Pacific Ocean--Nino 1+2, Nino 3, Nino 3.4, and Nino 4--as candidate predictors (Table 3.8).

(4) Monthly search query data from Google Trends [142]. Google trends data is an unbiased sample of Google search data, and only monthly-scale data is provided for time beyond five years. Each data point is normalized by the highest search volume of the term within the geography and time range. We define 41 search terms in Spanish relevance to the symptoms, biology and treatment of dengue (Table 3.9). For each term, we retrieve its monthly search popularity from 2004 to present on Google Trends. Next we apply a cubic spline method to disaggregate the monthly popularity to weekly scale, and negative values are set to 0 and values larger than 100 are set to 100 [83]. For each

study area, if a search term was searched by very few people (no time series is returned), we exclude the term as predictors.

3.4 RESULTS

3.4.1 Performance of early detection systems in each study area

We first selected optimal combinations of predictor time series within each data category and also across all data categories using cross-validation on training period. Regardless of locations, the MEWMA-FFS framework selects different sequential combinations of time series within the same data category for a range of values of penalty parameter ω (Figure 3.4), which indicates that ω does affect the selection of time series by adding penalty to false alarms. When optimizing time series across all data categories (*mix system)*, the framework trends to select a mixture of time series from different data categories as optimal combinations. The top one time series selected within each early detection system is not usually influenced by different penalty values, and optimal combinations are chosen from a small subset of time series in corresponding data categories.

In Mexico, the optimal combination of predictors within *incidence system* (public health surveillance data) includes *weekly incidence* and *weekly incidence (slope)* with a penalty parameter of 1 (Figure 3.4). In out-of-sample evaluation, the system sounds alarms for large dengue waves around the middle within the detection window of a corresponding outbreak (Table 3.3, Figure 3.4). The system also sound false alarms during 2015-2016 and 2016-2017 periods when the number of dengue cases did not meet the criteria of large outbreaks. The ‘winner’ *climate system* is the one with penalty parameter 2 and selects only one predictor time series -- *average temperature* (Table 3.3, Table 3.4, Figure 3.1). It detects the 2013-2014 and 2014-2015 outbreaks earlier than the *incidence system* with no false alarms triggered. The *Google Trends (GT) system* and *mix system* produce similar alarms with climate system in terms of alarm timing on average (Table 3.3, Figure 3.5). However, both systems trigger one false alarm and *GT* system also misses one large outbreak. The winner early detection system *climate system* also outperform all three baseline models in terms of alarm timing, precision and recall (Table 3.3, Figures 3.1 and 3.5). Even though *week-based trigger* sounds alarms for large outbreaks earlier than the winner *climate system*, it sounds alarms even for small outbreaks. We also compare performances of all early detection systems between cross-validation on training period (feature selection procedure) and out-of-sample evaluation on testing period to ensure consistent performance (Table 3.3, Figures 3.1 and 3.5). Across all models for Mexico, the performance are similar between training and testing period, which indicates the feature selection procedure has no potential overfitting issue. The result for Mexico shows that a subset of climate-relevance time series is the best option for early detection of large dengue outbreaks in Mexico.

Table 3.3: Performance of different early detection systems across cross validation and out-of-sample evaluation in Mexico

System	ω	Cross validation			Out-of-sample evaluation		
		Timing	Precision	Recall	Timing	Precision	Recall
Week-based trigger	-	0.94 (0.8-1.0)	0.67	1.0	0.92 (0.85-1.0)	0.5	1.0
Rise-based trigger	0.2	0.62 (0.47-0.8)	0.67	1.0	0.53 (0.46-0.6)	0.4	1.0
Univariate-incidence	0.6	0.42 (0.15-0.68)	0.4	0.67	0.52 (0.50-0.54)	0.67	1.0
Incidence	1	0.66 (0.45-0.93)	0.31	0.67	0.57 (0.54-0.60)	0.5	1.0
Climate	**2**	**0.90 (0.8-0.95)**	**0.86**	**1.0**	**0.86 (0.77-0.95)**	**1.0**	**1.0**
Google Trends	0.6	0.84 (0.46-1.0)	0.75	1.0	1.0 (1.0-1.0)	0.5	0.5
Mix	2	0.92 (0.84-0.95)	0.83	0.83	0.86 (0.77-0.95)	0.67	1.0

Table 3.4: Predictor time series included in ‘winner’ systems for each location

Location		
Mexico	San Juan	Iquitos
Average temperature	Average temperature Nino2 SSTA Nino3 SST (*acceleration)*	Cases Cases (*acceleration*) Cases (*slope*)

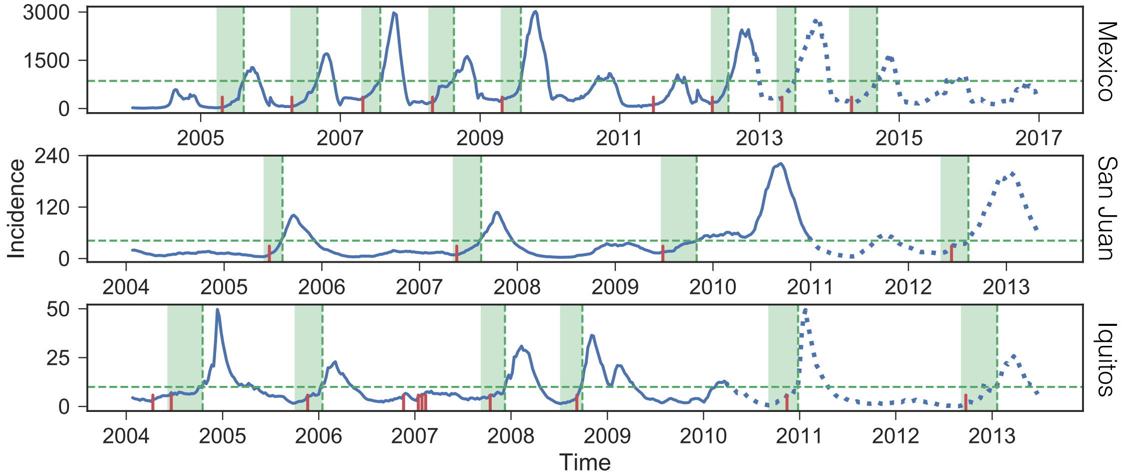

Figure 3.1: Performance of early detection systems for detecting large dengue waves for three study areas in cross validation and out-of-sample evaluation periods. Using weekly dengue incidence data as target time series (blue curves are cross validation periods, and blue dot curves are out-of-sample evaluation period), early detection systems were optimized to sound alarms as early as possible within pre-defined detection windows (shaded green area). Systems are selected to detect events defined by a threshold of 75% percentile of historical reported dengue cases (green horizontal and vertical dashed lines). Red bars indicate alarms triggered by the 'winner' system for each study area: *climate system* with $\omega = 2$ for Mexico; *climate system* with $\omega = 0.6$ for San Juan; *incidence system* with $\omega = 0.6$ for Iquitos, where ω is a penalty parameter to false alarms).

In San Juan, the *incidence system* with a penalty parameter of 1.0, including *dengue incidence* and *dengue incidence (slope)*, detects all large outbreaks within cross validation period, with an average of alarm timing of 0.28 and a precision of 0.75 (Table 3.5, Figure 3.4). In out-of-sample testing, it sounds an alarm for the only 2012-2013 outbreak earlier than any alarms for outbreaks within cross validation period, and produces one false alarm (Table 3.5, Figure 3.6). The climate sy*stem*, including three time series -- *average temperature*, *Nino3 SSTA* and *Nino SST (acceleration)*, performs the best across all models (both baseline models and other early detection systems) in terms of alarm timing, precision and recall (Figures 3.1 and 3.4, Tables 3.4 and 3.5).

Performance of the *mix system* is slightly lower than that of the *climate system*, and also sounds no false alarms. In out-of-sample evaluation, the performance of *climate* and *mix* system are also similar with each other on the single outbreak, both of which fall within the range of alarm timing in cross validation, respectively (Table 3.5, Figures 3.1 and 3.6). Surprisingly, even though *GT system* triggers false alarms and also misses one large outbreak in cross validation, it sounds an alarm for the single outbreak in out-of-sample testing periods as early as the *mix system*, and produces no false alarms (Table 3.4, Figure 3.6). Overall, performance of the *climate system* is more robust than any other systems for San Juan area across cross validation and out-of-sample testing period.

Table 3.5: Performance of different early detection systems across cross validation and out-of-sample evaluation in San Juan

System	ω	Cross validation			Out-of-sample evaluation		
		Timing	Precision	Recall	Timing	Precision	Recall
Week-based trigger	-	0.71 (0.53-1.0)	0.5	1.0	0.5 (0.5-0.5)	0.5	1.0
Rise-based trigger	0.2	0.56 (0.4-0.74)	0.6	1.0	0.64 (0.64-0.64)	0.5	1.0
Univariate-incidence	1.4	0.32 (0.07-0.58)	0.6	1.0	0.79 (0.79-0.79)	0.5	1.0
Incidence	1.0	0.28 (0.07-0.58)	0.75	1.0	0.64 (0.64-0.64)	0.5	1.0
Climate	0.6	**0.84** **(0.60-0.95)**	**1.0**	**1.0**	**0.64** **(0.64-0.64)**	**1.0**	**1.0**
Google Trends	4	0.78 (0.63-0.93)	0.67	0.67	0.86 (0.86-0.86)	1.0	1.0
Mix	4	0.55 (0.1-0.93)	1.0	1.0	0.86 (0.86-0.86)	1.0	1.0

When detecting large dengue outbreaks for Iquitos metropolitan area in Peru, the *incidence* system, including *dengue incidence, dengue incidence (acceleration) and dengue incidence (slope)*, triggers alarms for large outbreaks around the middle of the detection windows on average, with a precision of 0.44, which is slightly better than the univariant version of MEWMA baseline with only *dengue incidence* time series (Table 3.6, Figures 3.1, 3.4 and 3.7). It indicates that slope and acceleration of dengue incidence time series increase the performance marginally. In out-of-sample evaluation, the system does not trigger any false alarms and sounds alarms for large waves with an average alarm timing of 0.64 (Table 3.6, Figure 3.1). We found that during the 2006-2007 period, the incidence time series is more noisy than that within the 2012-2013 period, which might explain the multiple false alarms triggered in cross validation. All other three early detection systems (*climate, GT,* and *mix systems*) are able to detect all large outbreaks within cross validation period, and outperform the *incidence* system in terms of alarm timing and precision (Table 3.6, Figure 3.7). In out-of-sample evaluation, the alarm timing of *climate system* is consistent with that in cross validation, while the *mix system* sounds alarms for large outbreaks later than the lower boundary of alarm timing in cross validation and the *GT system* fails to detect any large outbreaks (Table 3.6, Figure 3.7). This phenomenon reveals potential overfitting of the models. Optimal combinations of predictor time series in both *mix* and *GT* system include search query data where we suspect the issue stems from. Crowd-source data, such as search popularity of Google terms, are affected easily by social media and sentiment of human communities. Similar issues have been found in previous studies [143,144].

Table 3.6: Performance of different early detection systems across cross validation and out-of-sample evaluation for Iquitos

System	ω	Cross validation			Out-of-sample evaluation		
		Timing	Precision	Recall	Timing	Precision	Recall
Week-based trigger	-	0.48 (0-1.0)	0.67	1.0	0.80 (0.75-0.84)	0.67	1.0
Rise-based trigger	0.4	0.74 (0.33-0.89)	0.4	1.0	0.79 (0.63-0.95)	0.5	1.0
Univariate-incidence	1	0.49 (0.17-0.89)	0.4	1.0	0.40 (0.38-0.42)	1.0	1.0
Incidence	0.6	0.57 (0.25-0.89)	0.44	1.0	**0.63** **(0.38-0.89)**	**1.0**	**1.0**
Climate	0.6	0.89 (0.58-1.0)	0.57	1.0	0.88 (0.88-0.88)	0.5	0.5
Google Trends	0.6	0.92 (0.83-1.0)	0.44	1.0	-	-	-
Mix	1	**0.67** **(0.54-0.75)**	**0.8**	**1.0**	0.23 (0.19-0.26)	1.0	1.0

3.4.2 Performance of area-specific early detection systems on other areas

To study if an optimal early detection system for one specific location has predictive power for other locations, we use predictor time series selected by the 'winner' early detection system for each of the three study areas (Table 3.4) to detect large dengue waves in the other two locations. When detecting large dengue outbreaks for Mexico, the San Juan-based early detection system sound alarms for large dengue outbreaks as early as that of the Mexico 'winner' system, however, it produces more false alarms and misses more than 75% of all large outbreaks (Figure 3.2). The Iquitos-based system does not outperform the Mexico 'winner' system in terms of all three metrics. For San Juan, the three early detection systems detect all large outbreaks occurred, however, the Mexico-

and Iquitos- based systems sound alarms later and the Iquitos-based system produces more false alarms than the San Juan ‘winner’ system (Figure 3.2). Interestingly, in Iquitos, San Juan-based system triggers alarms for large dengue outbreaks earlier than the Iquitos ‘winner’ system, while sounds more false alarms. Mexico-based early detection system has similar alarm timing with the Iquitos ‘winner’ system and produces less false alarms, but it fails to detect more than half of the large outbreaks for Iquitos (Figure 3.2). It indicates that early detection systems optimized for one location can only provide limited information for early detection of dengue outbreaks in other locations and is prone to sound more false alarms.

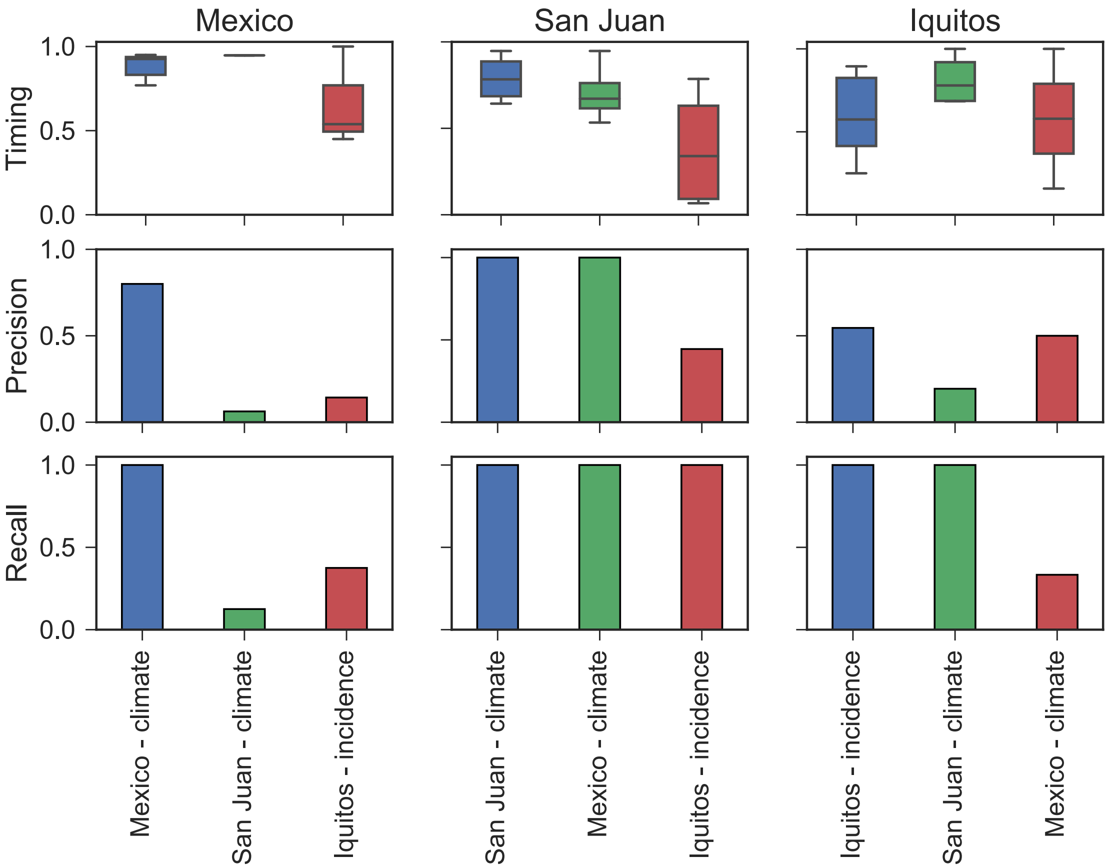

Figure 3.2: Performance of each local 'winner' early detection system on detecting large dengue waves in other study areas, in terms of alarm timing relative to the size of corresponding detection window, proportion of alarms indicating actual large outbreaks (precision) and proportion of outbreaks detected (recall). Evaluations are performed using leave-1-year-out cross validation on entire time periods available for each study area. Alarm timing is computed by averaging alarm timing across all large outbreaks in each area.

3.5 DISCUSSION

Over the past years, we have developed a hierarchical framework MEWMA-FFS to detect emerging and re-emerging infectious disease outbreaks. By applying this two-layer framework, we are able to not only build models for early detection of infectious disease outbreaks but also choose an optimized subset of predictors for the model from thousands of candidate predictor data sources. In this study, we use the MEWMA-FFS framework in detecting large dengue outbreaks to three dengue-endemic areas -- Mexico, San Juan in Puerto Rico and Iquitos in Peru. To identify large outbreaks from small ones, We introduce a parameter *baseline threshold* to label only observations between baseline threshold and event threshold as non-outbreak. Based on thirteen years of data, the framework identifies optimal combinations of predictor time series for each location, and the combinations vary between locations – both Mexico and San Juan system are driven by time series from climate stations, and Iquitos model is dominated by time series from public health surveillance systems. With optimized predictor time series, early detection systems are able to detect large dengue outbreaks earlier and produce less false alarms than baseline models including only one time series *dengue incidence*. We also found that predictor time series optimized for one specific location may not be applied to other locations directly for the purpose of early detection of large dengue outbreaks.

Results in the study are consistent with previous research. Dengue forecasting models including climate data, such as temperature, precipitation, relative humidity, and sea surface temperature, usually achieve good performance in regions along coasts

[133,134]. We speculate this is because areas along coasts are influenced by El Niño–Southern Oscillation (ENSO) the most and ENSO has a directed effect on climate change and disrupts normal weather patterns [145]. The differences between the optimal combinations for San Juan and Mexico may result from geographical heterogeneity which causes various dynamics of dengue transmission [146]. On the other hand, Iquitos is far away from the Pacific Ocean border and located along the Amazon River. It experiences a tropical rainforest climate. That is, there is constant rainfall throughout the year without a distinct dry season, and the temperature usually ranges from 21 to 33 degree Celsius [147]. It indicates the climate does not vary much throughout the year and thus are not good predictors for large dengue outbreaks. In addition, dengue incidence in Iquitos is very low, even during the peak of a large outbreak (~50 cases), compared to that in Mexico and San Juan, which may be another reason that makes the climate-related predictors do not stand out over public health surveillance data. Google search terms relevance to dengue, which have been applied to dengue nowcasting in multiple areas [44,83,84], never beat other predictor time series in early detection of large dengue outbreaks in any of the three locations. There are two potential reasons: (1) Google Trends reflects the search popularity of terms relevance to dengue infection, and the suddenly increase in search popularity may indicate already increased dengue incidence in a population. This is usually later than the effect of climate factors, which influence the biology of mosquitoes--the vector transmitting dengue virus between humans. (2) Search popularity data retrieved from Google Trends is only available in monthly scale, and thus

is disaggregated into weekly scale in the study. Applying any methods to disaggregate data may smooth or change the actual trends in time series.

We emphasize that the conclusion in this study should not be applied to other dengue forecasting models with different purpose directly. The results here are based on the MEWMA model and forward feature selection algorithm to detect large dengue outbreaks that is defined by 75th percentile of the distribution of weekly dengue incidence in historical data. Studies have shown that climate related predictors did not improve dengue short-term forecasting significantly using a SARIMA framework for Mexico [65], and that a Gaussian process (GP) regression, ignoring environmental factors, outperformed other models in the Dengue Forecasting Project [148]. In addition, even for the same location with the same MEWMA-FFS framework, choosing an alternative target time series, such as the number of laboratory-confirmed dengue cases or dengue hemorrhagic fever, may result in different optimal combinations of predictor time series.

Our study is also limited to no more than thirteen years of historical data for each location as Google Trends is only available since 2004. Even though we designed our study carefully by selecting predictor time series using cross validation on training period and evaluating optimal combinations of predictors on testing period, we cannot guarantee the conclusion would be the same when more data is available in the near future. Our previous study has shown that the length of time series does affect feature selection and the performance of systems [110]. Therefore, we recommend to evaluate constructed early detection systems periodically. In addition, a common issue for research in emerging and re-emerging outbreaks is that there is no sufficient historical data available since such outbreaks are very sparse. In our study, the out-of-sample evaluation is only based on no more than four years of data, which may limit the reliability of the optimal combinations of predictor time series. In future studies, we need to develop algorithms for simulating plausible emerging or re-emerging outbreaks [55], and use the bootstrapping method [149] to sample outbreaks with replacement for generating time series containing more outbreaks. With more data available, we can improve the model validation process.

While we have already included more than one hundred candidate predictor time series in the study, considering other predictors may provide more information to benefit the early detection of large dengue outbreaks. Electronic Health Records data, which is useful in influenza forecasting and nowcasting [37,56,93,102,150], may be informative for dengue early detection as well. Numerous Internet-sourced data has been validated and applied to infectious diseases surveillance besides Google Trends, such as Twitter, Wikipedia, HealthMap etc. Such data should be considered if available in target areas. However, we should take extra caution when applying those Internet-sourced data, since they are susceptible to public events such as media news, public sensibility and fear to infectious diseases. In addition, some other environmental factors are proved to be correlated with dengue infection, such as absolute humidity and vegetation [151].

Building models to detect emerging and re-emerging infectious disease outbreaks is always an important task for the public health community. It can help public health agencies to detect potential outbreaks in advance, and thus have enough lead time to implement effective intervention strategies and distribute vaccines and antiviral drugs. Currently, researchers mainly focus on building models for specific locations, which may not be applied to other locations directly as shown in this study. It is still a big challenge for building early detection systems for locations lacking of established surveillance systems or experiencing political unrest. Moreover, for a single region, data relevance to emerging and re-emerging outbreaks is often very limited considering the relative low frequency of those events. Therefore, it is vital to build universal early detection models to cover multiple regions. Multiple ways can be applied to achieve this goal, such as grouping locations based on their similarity in population, climate and geography etc., and one model can be built for each cluster of locations by combining their candidate time series. With this, it not only benefits regions without high-quality public health data but also increases statistical power of early detection models.

3.6 SUPPLEMENTAL INFORMATION

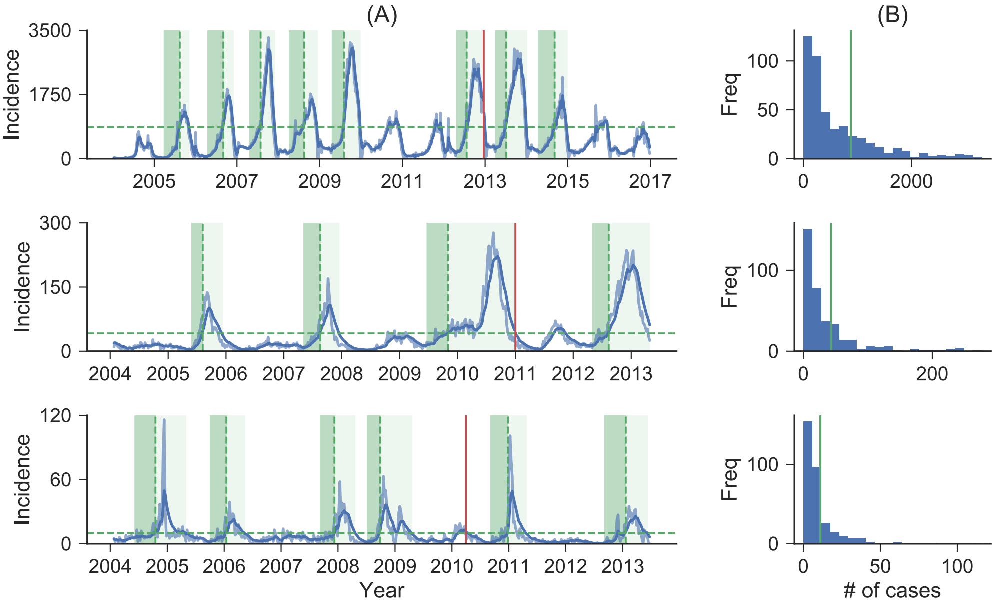

Figure 3.3: Historical dengue incidence in the country of Mexico (panel 1), San Juan metropolitan area of Puerto Rico (panel 2), and Iquitos metropolitan area of Peru (panel 3). (A) Time series of raw and smoothed number of dengue cases. The light blue curve is the actual weekly number of cases, and the dark blue curve is the smoothed time series. The horizontal green dash line represents the event threshold, and vertical green dash line indicates the start of each event. Dark green shadow areas are detection windows prior to the start of events, and alarms triggered within these areas are true alarms. Light green shadow areas are periods experiencing large dengue outbreaks, and alarms falling in these periods are neutral alarms. Alarms outside of the two areas are false alarms. Data before the vertical red line are used for model training and others are for out-of-sample evaluation. (B) Distribution of dengue incidence across training period. The vertical green line shows 75th percentile of the distribution.

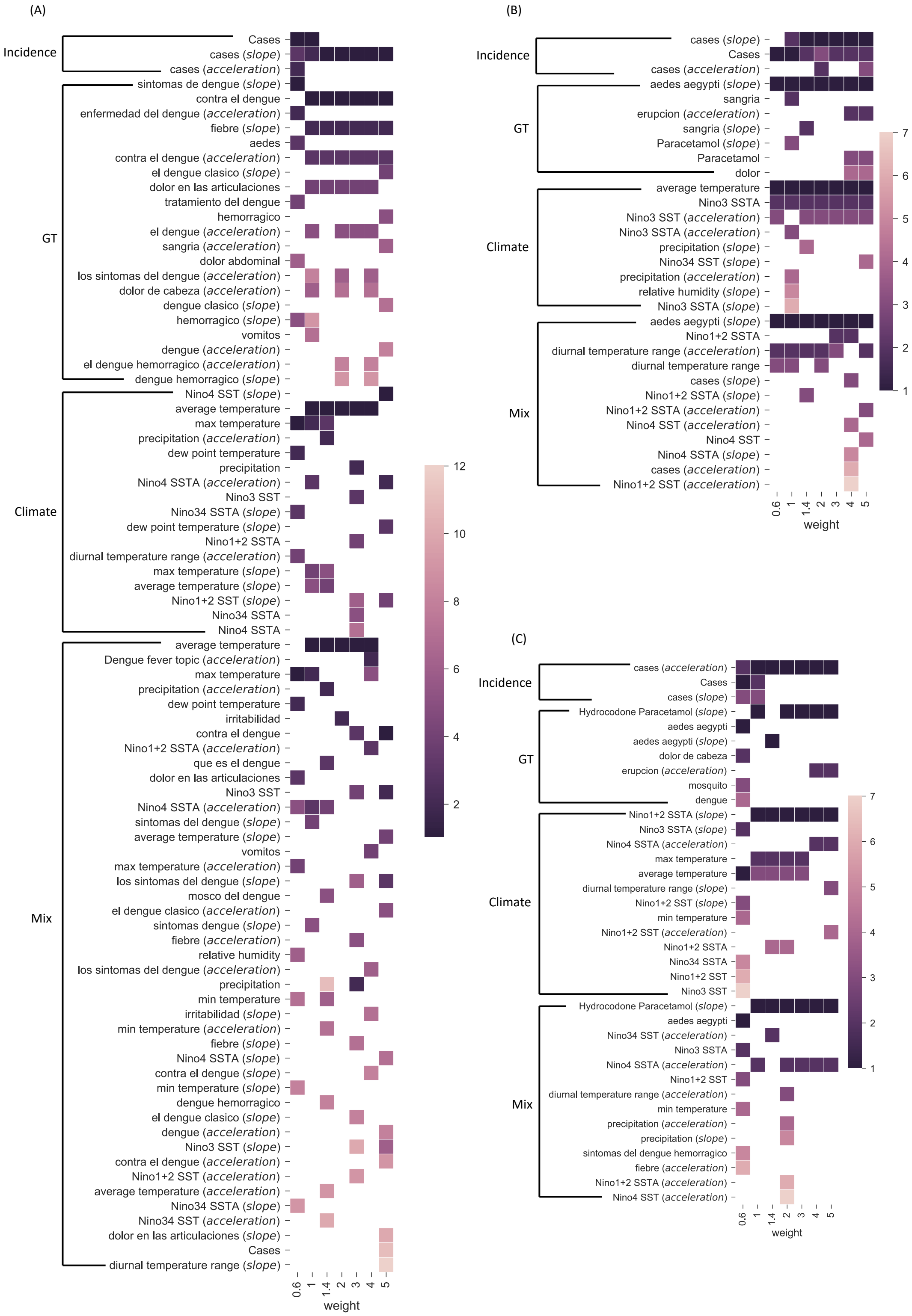
(A)
Incidence
GT
Climate
Mix
Cases
cases (*slope*)
cases (*acceleration*)
sintomas de dengue (*slope*)
contra el dengue
enfermedad del dengue (*acceleration*)
fiebre (*slope*)
aedes
contra el dengue (*acceleration*)
el dengue clasico (*slope*)
dolor en las articulaciones
tratamiento del dengue
hemorragico
el dengue (*acceleration*)
sangria (*acceleration*)
dolor abdominal
los sintomas del dengue (*acceleration*)
dolor de cabeza (*acceleration*)
dengue clasico (*slope*)
hemorragico (*slope*)
vomitos
dengue (*acceleration*)
el dengue hemorragico (*acceleration*)
dengue hemorragico (*slope*)
Nino4 SST (*slope*)
average temperature
max temperature
precipitation (*acceleration*)
dew point temperature
precipitation
Nino4 SSTA (*acceleration*)
Nino3 SST
Nino34 SSTA (*slope*)
dew point temperature (*slope*)
Nino1+2 SSTA
diurnal temperature range (*acceleration*)
max temperature (*slope*)
average temperature (*slope*)
Nino1+2 SST (*slope*)
Nino34 SSTA
Nino4 SSTA
average temperature
Dengue fever topic (*acceleration*)
max temperature
precipitation (*acceleration*)
dew point temperature
irritabilidad
contra el dengue
Nino1+2 SSTA (*acceleration*)
que es el dengue
dolor en las articulaciones
Nino3 SST
Nino4 SSTA (*acceleration*)
sintomas del dengue (*slope*)
average temperature (*slope*)
vomitos
max temperature (*acceleration*)
los sintomas del dengue (*slope*)
mosco del dengue
el dengue clasico (*acceleration*)
sintomas dengue (*slope*)
fiebre (*acceleration*)
relative humidity
los sintomas del dengue (*acceleration*)
precipitation
min temperature
irritabilidad (*slope*)
min temperature (*acceleration*)
fiebre (*slope*)
Nino4 SSTA (*slope*)
contra el dengue (*slope*)
min temperature (*slope*)
dengue hemorragico
el dengue clasico (*slope*)
dengue (*acceleration*)
Nino3 SST (*slope*)
contra el dengue (*acceleration*)
Nino1+2 SST (*acceleration*)
average temperature (*acceleration*)
Nino34 SSTA (*slope*)
Nino34 SST (*acceleration*)
dolor en las articulaciones (*slope*)
Cases
diurnal temperature range (*slope*)
0.6
1
1.4
2
3
4
5
weight
12
10
8
6
4
2
(B)
Incidence
GT
Climate
Mix
cases (*slope*)
Cases
cases (*acceleration*)
aedes aegypti (*slope*)
sangria
erupcion (*acceleration*)
sangria (*slope*)
Paracetamol (*slope*)
Paracetamol
dolor
average temperature
Nino3 SSTA
Nino3 SST (*acceleration*)
Nino3 SSTA (*acceleration*)
precipitation (*slope*)
Nino34 SST (*slope*)
precipitation (*acceleration*)
relative humidity (*slope*)
Nino3 SSTA (*slope*)
aedes aegypti (*slope*)
Nino1+2 SSTA
diurnal temperature range (*acceleration*)
diurnal temperature range
cases (*slope*)
Nino1+2 SSTA (*slope*)
Nino1+2 SSTA (*acceleration*)
Nino4 SST (*acceleration*)
Nino4 SST
Nino4 SSTA (*slope*)
cases (*acceleration*)
Nino1+2 SST (*acceleration*)
0.6
1
1.4
2
3
4
5
weight
7
6
5
4
3
2
1
(C)
Incidence
GT
Climate
Mix
cases (*acceleration*)
Cases
cases (*slope*)
Hydrocodone Paracetamol (*slope*)
aedes aegypti
aedes aegypti (*slope*)
dolor de cabeza
erupcion (*acceleration*)
mosquito
dengue
Nino1+2 SSTA (*slope*)
Nino3 SSTA (*slope*)
Nino4 SSTA (*acceleration*)
max temperature
average temperature
diurnal temperature range (*slope*)
Nino1+2 SST (*slope*)
min temperature
Nino1+2 SST (*acceleration*)
Nino1+2 SSTA
Nino34 SSTA
Nino1+2 SST
Nino3 SST
Hydrocodone Paracetamol (*slope*)
aedes aegypti
Nino34 SST (*acceleration*)
Nino3 SSTA
Nino4 SSTA (*acceleration*)
Nino1+2 SST
diurnal temperature range (*acceleration*)
min temperature
precipitation (*acceleration*)
precipitation (*slope*)
sintomas del dengue hemorragico
fiebre (*acceleration*)
Nino1+2 SSTA (*acceleration*)
Nino4 SST (*acceleration*)
0.6
1
1.4
2
3
4
5
weight
7
6
5
4
3
2
1

Figure 3.4

Figure 3.4: Predictor time series selected by early detection systems with different penalty weight ω to false alarms. Systems were optimized within each data category (incidence, climate, Google Trends(GT)) and across all data categories. Darkness of the color indicates the sequence of time series selected. The lighter the color is, the later the corresponding time series is selected in a system. The time series corresponding to white color represents it is never selected in systems.

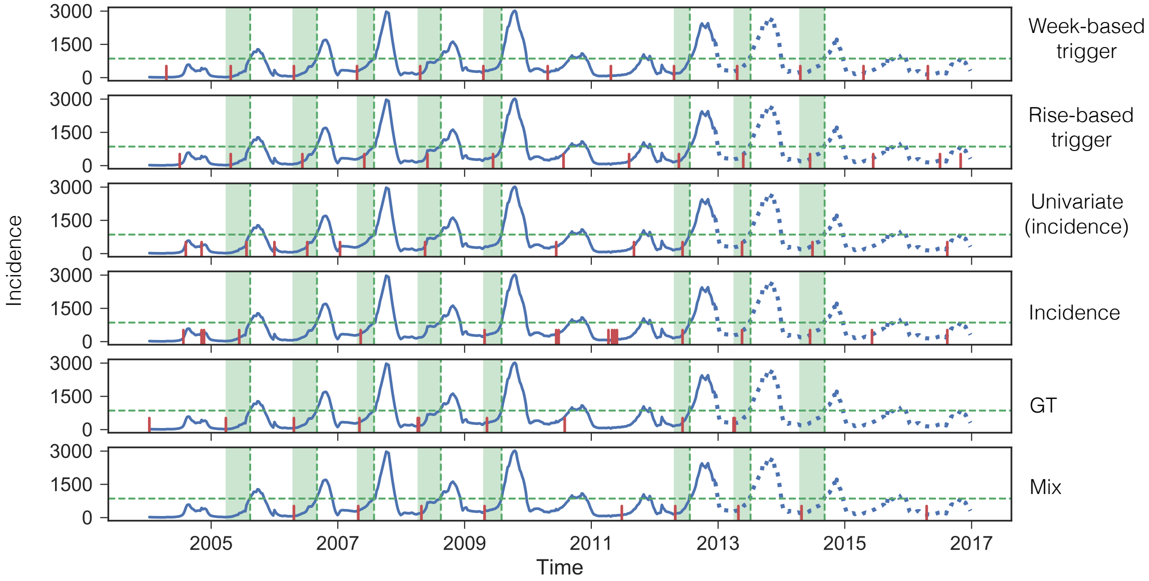

Figure 3.5: Alarms triggered across various early detection systems for Mexico in detecting large dengue outbreaks. Using weekly dengue incidence data as target time series (blue curves are cross validation periods, and blue dot curves are out-of-sample evaluation period), early detection systems were optimized to sound alarms as early as possible within pre-defined detection windows (shaded green area). Predictor time series were selected to detect events defined by a threshold of 75% percentile of the distribution of historical dengue cases (green horizontal and vertical dashed lines). Red bars indicate alarms triggered by the corresponding baseline or early detection system.

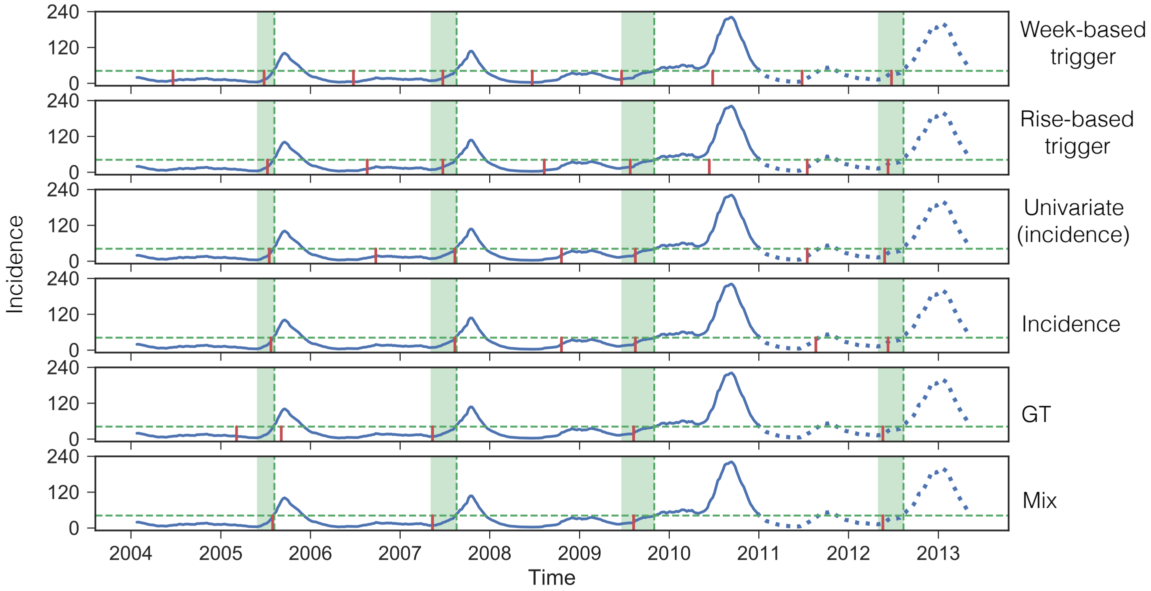

Figure 3.6: Alarms triggered across various early detection systems for San Juan in detecting large dengue outbreaks. Using weekly dengue incidence data as target time series (blue curves are cross validation periods, and blue dot curves are out-of-sample evaluation period), early detection systems were optimized to sound alarms as early as possible within pre-defined detection windows (shaded green area). Predictor time series were selected to detect events defined by a threshold of 75% percentile of the distribution of historical dengue cases (green horizontal and vertical dashed lines). Red bars indicate alarms triggered by the corresponding baseline or early detection system.

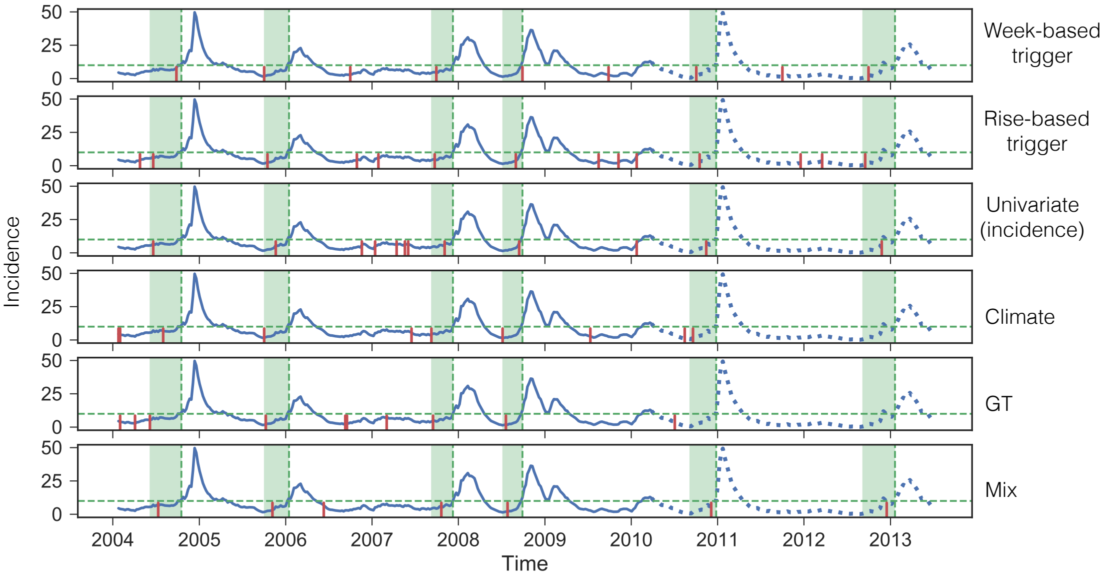

Figure 3.7: Alarms triggered across various early detection systems for Iquitos in detecting large dengue outbreaks. Using weekly dengue incidence data as target time series (blue curves are cross validation periods, and blue dot curves are out-of-sample evaluation period), early detection systems were optimized to sound alarms as early as possible within pre-defined detection windows (shaded green area). Predictor time series were selected to detect events defined by a threshold of 75% percentile of the distribution of historical dengue cases (green horizontal and vertical dashed lines). Red bars indicate alarms triggered by the corresponding baseline or early detection system.

Table 3.7: Training and testing periods for each study area

Region	Training period	Testing period
Mexico	01/11/2004 – 02/20/2013	02/21/2013 – 01/01/2017
San Juan	01/25/2004 – 04/29/2011	04/30/2011 – 04/29/2013
Iquitos	01/25/2004 – 07/01/2010	07/02/2010 – 06/30/2013

Table 3.8: Target and candidate predictor time series from each study area

Region	Date range	Data category	Predictors [c]
Mexico	2004 - 2017, weekly	Incidence [a]	1 × 3
		Climate	7 × 3
		Sea surface temperature	8 × 3
		Google Trends [b]	40 × 3
San Juan	2004 - 2013, weekly	Incidence [a]	1 × 3
		Climate	7 × 3
		Sea surface temperature	8 × 3
		Google Trends [b]	27 × 3
Iquitos	2004 - 2013, weekly	Incidence [a]	1 × 3
		Climate	7 × 3
		Sea surface temperature	8 × 3
		Google Trends [b]	37 × 3

Table notes: a. Target time series; b. Search terms included are shown in Table 3.9; c. The first number in each row indicates the total number of predictor time series, and the second number shows the three levels per time series.

Table 3.9: Candidate Google search terms. '×' indicates that the search term is not included as a candidate predictor in the corresponding area.

Search term	Mexico	San Juan	Iquitos
dengue			
el dengue			
Que es el dengue			
Dengue clasico		×	
El dengue clasico		×	×
Tipos de dengue			
Casos de dengue	×	×	
Enfermedad del dengue		×	
hemorragico			
el dengue hemorragico		×	×
dengue hemorragico			
sintomas del dengue			
sintomas de dengue			
sintomas dengue			
sintomas del dengue hemorragico		×	
los sintomas del dengue		×	
mosco del dengue		×	×
mosquito del dengue			
tratamiento para el dengue		×	
contra el dengue		×	×
tratamiento del dengue		×	
tratamiento dengue		×	
dengue fever			
virus del dengue		×	
Mosquito			
Aedes			
aedes aegypti			
Fiebre			
Dolor			
dolor de Cabeza			
Erupcion			
Sangria			
dolor abdominal			
dolor en las articulaciones		×	
Vomitos			
Hematomas			
Somnolencia			
Irritabilidad			
Paracetamol			
Hydrocodone / Paracetamol			
Oxycodone / Paracetamol			

Chapter 4: The effects of population structures on infectious disease dynamics and prediction

4.1 ABSTRACT

Infectious diseases are transmitted via contacts between individuals which form a contact network. Network structure can be complex, including heterogeneity in the number of contacts, clustering of contacts within coherent subpopulations, and asymmetric contacts in which one individual is likely to infect another, but the reverse is not true. For example, health care workers (HCWs) in a hospital setting may frequently contact infectious patients. Ignoring such population structures can bias epidemic predictions, as occurred during the 2014 Ebola epidemic in West Africa. Here, we explore the interactive effects of heterogeneous, clustered, and directed contacts on the unfolding of an epidemic. Using a low-dimension system of ordinary differential equations, we find that heterogeneous and directed contacts significantly impact the timing and magnitude of spread, while clustering has a relatively minor effect. Using simulated data collected in early phase of an epidemic, we further assess the ability of various models to infer transmission rates and make predictions based on data from an emerging outbreak. If we ignore all three network structures, our models overestimate total incidence and the timing and magnitude of highest incidence (i.e., the epidemic peak) by more than 10%, 6 days, and 20% respectively. By incorporating heterogeneity,

we reduce these errors to 5%, 3 days and 0% respectively; by incorporating heterogeneity, clustering and directed contacts, the error nearly disappears.

4.2 INTRODUCTION

Ebola virus, a lethal human pathogen, caused the largest known Ebola epidemic in West Africa during 2014-2015 since it first appeared. The epidemic originated from Guinea in December 2013, and rapidly expanded to three other countries, including Libera, Sierra Leona, and Nigeria. Imported cases were also reported in countries outside West Africa. As of February 2016, the total number of probable, suspected and confirmed Ebola cases in the epidemic was 28,639 with 11,316 deaths globally which exceeded the total number of cases and deaths from all previous outbreaks [152]. To support public health agencies for disease-control efforts, epidemiologists developed and parameterized mathematical models to predict the epidemic trajectory. By September 2015, 15 publications (including 22 models) provided numerical forecasts of cumulative Ebola incidence [153]. However, 18 of the 22 models, which assumed exponential growth in the initial phase of the epidemic, overestimated the future number of cases. For example, using the EbolaResponse modeling tool developed by Center for Disease Control and Prevention (CDC), the estimated total number of cases in Liberia and Sierra Leone would be 550,000 by January 20, 2015 [154].

The discrepancy between prediction and actual cumulative incidence might be caused by early and effective intervention. However, even during the early phase of the epidemic when public health intervention has not been implemented broadly, the epidemic dynamic in district level was largely characterized by sub-exponential, instead of exponential, growth patterns [155]. The departure from the exponential assumption of the *mass action* compartmental model could affect the estimation of final epidemic size, since the effective reproduction number (R_0) declines rapidly for sub-exponential growth within the first three to five disease generations while does not change in exponential growth [156]. This phenomena might stem from clustering in the population, heterogeneous mixing, spatial effects and reactive behavior changes *et al*. A study shows that predictive models including decay terms, heterogeneous contact patterns or other methods to constrain incidence growth tend to have lower forecast error [153]. An agent-based simulation model with various population structures (Ebola treatment units, households *et al*.), was also able to replicate the sub-exponential growth patterns [157]. Scarpino *et al*. has confirmed that clustered transmission did exist in the population by analyzing Ebola virus genomic and epidemiological data from Sierra Leone [158]. In addition, during the Ebola epidemic, at least 10 clusters of Ebola cases among health care workers (HCWs) at non-Ebola treatment units have been reported [159,160], which were initialized by patients who infected HCWs by seeking medical attentions. Unlike the transmission within households, the transmission between patients and HCWs is asymmetric, with patients more likely to infected HCWs than to be infected by HCWs [161]. HCWs have accounted for up to 25% of Ebola cases during previous Ebola outbreaks [162], which indicates the importance of including health care settings in modeling Ebola epidemics.

The impact of various population structures on epidemics has attracted attentions of mathematical epidemiologists. Researchers usually use network models to study the diverse interactions underlying the spread of infectious diseases via either analytical methods or simulations [163–166]. In a network model, nodes represent individuals in a population and edges indicates the interaction between two individuals. Degree distribution describes the distribution of the number of contacts per individual. A previous theoretical study shows that for the same R_0, the total incidence of the epidemic could be different on networks with different degree distributions [163]. Bansal *et al* suggests that human contact patterns are more heterogeneous than assumed by homogeneous-mixing models, and that mass action model is not appropriate for populations with heterogeneous contact patterns [164]. In another study, Meyers *et al* investigated the impact of hospital-based transmission on the fate of an epidemic [161]. They used directed edges starting from average people to HCWs to reveal the disease transmission between patients and HCWs within a network (meaning an average person is more likely to infect HCWs by seeking medical attention than to be infected by HCWs), and undirected edges to represent the transmission between two average individuals. They find that the probability of an epidemic and the expected fraction of a population infected during an epidemic can be different when considering the hospital-based transmission. Volz *et al* studied the joint impact of clustering and heterogeneity in contact patterns on epidemics [167]. It shows the interaction between clustering and heterogeneity is complex, and clustering always slows down an epidemic while increasing clustering and heterogeneity simultaneously can decrease final epidemic size.

With different population structures being studied, epidemiological models become more and more complex, however, there is little study on the contributions of those structures to epidemic dynamics, and the trade-off between model complexity and model performance on prediction. For example, researchers usually fit a mathematical model to the data collected from an epidemic to estimate transmission parameters, and then project the total incidence to future. In some cases, models without considering all population structures might achieve reasonable performance by adjusting transmission parameters, even though the estimated parameters are not consistent with ground truth.

In light of the Ebola epidemic in 2014, we developed deterministic Susceptible-Exposed-Infected-Recovered (SEIR) Ordinary Differential Equations (ODEs) to model the spread of infectious diseases on static networks, where heterogeneous contact pattern, clustering and directed contacts are considered, based on the edge-based compartmental modelling approach [167–170]. Using the network-based SEIR model, we first investigated contributions of different population structures on epidemic dynamics with various combinations of network parameters. We find that both heterogeneous and directed contacts contribute to the epidemic dynamics significantly, and the effect of clustering is relatively small. Next, we explore the ability of different models to make predictions using data from a simulated epidemic. The result suggests that a model without any population structures always overestimate total incidence, magnitude and timing of the epidemic peak by more than 10%, 20% and 6 days respectively, while a network model with only heterogeneity can make better predictions that lowers the errors by 5%, 20% and 3 days. Our study provides not only theoretical understanding of contributions of different population structures on disease dynamics, but also insight into practical modeling and data collection suggestions to achieve better surveillance for future epidemics.

4.3 MATERIALS AND METHODS

4.3.1 SEIR network model derivation

Variables and parameters involved in the model are described in Table 4.1. We consider a susceptible-exposed-infected-recovered (SEIR) model on a static network, in which there are three different population structures—heterogeneous, directed and clustered contacts. Heterogeneous contacts are described by *undirected edges* between two individuals, meaning that transmission can occur in either direction. Each individual in the network has a different number of undirected edges. Clustered contacts are represented by *triangles*, each of which includes three individuals and edges. A triangle is explained as two friends of one individual are also friends. The disease transmission between two individuals within a triangle is bidirectional. Directed contacts are defined as *directed edges*, which point from average people to HCWs. *In-degree* and *out-degree* represent the number of directed edges incoming from and outgoing to other individuals, respectively. In a network, each individual is a member of a random number of in-degree, out-degree, undirected edges, and triangles. The *degree distribution* shows the probability that a randomly chosen individual will have a particular combination of in-degree, out-degree, undirected edges and triangles. The network structure is captured by a probability generating function (PGF):

$$\psi(x,y,z,v) = \sum_{k_i,k_o,k_u,k_c} p(k_i,k_o,k_u,k_c)x^{k_i}y^{k_o}z^{k_u}v^{k_c} \tag{4.1}$$

Initially, each individual in the network is susceptible until infected individuals are introduced. Individuals infected by others in infected state at a constant transmission rate β transit to exposed state. Individuals in exposed state are moved to infected state at a constant rate σ, and those in infected state are transmitted to recovered state at a constant rate γ. Once recovered, an individual cannot be infected anymore. We derive the SEIR model based on the approach of edge-based compartmental modelling. We define a *test individual* u, which is sampled randomly from a network. The probability that the individual u is in a given state is equal to the proportion of individuals in that state. We assume that the individual u does not transmit infection to its neighbors [168]. By this assumption, we can eliminate the dependence of the states between two neighbors of u since a neighbor of u can infect another neighbor of u by infection traveling through u otherwise. Based on the assumption, each neighbor of u can infect u independently. This does not affect the probability that u is in a given state, and thus has no impact on calculating the proportion of individuals in that state.

Table 4.1 Definitions of variables and parameters in the model

Parameter	Definition
σ	Exposure rate
β_u	Transmission rate per contact via undirected edges
β_d	Transmission rate per contact via directed edges
β_c	Transmission rate per contact on triangles
γ	Recovery rate
$p(k_i, k_o, k_u, k_c)$	The probability that an individual is a member of k_u undirected edges, k_i in-degree edges, k_o out-degree edges, and k_c triangles
$S(0)$	The proportion of individuals that are susceptible at time 0
S, E, I, R	The proportion of individuals in susceptible, exposed, infected or recovered states
$\phi_{S,u}, \phi_{E,u}, \phi_{I,u}, \phi_{R,u}$	The probability that a neighbor of u along an undirected edge is susceptible, exposed, infected or recovered, and has not transmitted infection to u given that it had not at time 0
$\phi_{S,d}, \phi_{E,d}, \phi_{I,d}, \phi_{R,d}$	The probability that a neighbor of u along a directed edge is susceptible, exposed, infected or recovered, and has not transmitted infection to u given that it had not at time 0
ϕ_{XY}	The probability that two neighbors of u in a same triangle are in states X and Y, and have not transmitted infection to u given that they had not at time 0 ($X, Y \in \{S, E, I, R\}$)
θ_u	The probability that a neighbor of u along an undirected edge has not transmitted infection to u given that it had not at time 0
θ_d	The probability that a neighbor of u along a directed edge has not transmitted infection to u given that it had not at time 0
θ_c	The probability that two neighbors of u in a same triangle have not transmitted infection to u given that they had not at time 0
$\psi(x, y, z, v)$	The probability generating function for generating the probability that an individual is a member of k_u undirected edges, k_i in-degree edges, k_o out-degree edges and k_c triangles
A	The rate that a neighbor of u in a triangle is infected by individuals outside the triangle

We start with calculating the proportion of individuals in susceptible state at time t $(S(t))$, which is equivalent to the probability that the individual u is in susceptible state. At time 0, infection is introduced into the network, and the proportion of susceptible individuals is $S(0)$. We assume that the individual u has k_u undirected edges, k_i in-degree edges, k_o out-degree edges and k_c triangles, and is susceptible at time 0. Then at time t, the probability that it is still susceptible is $S(0)\theta_d^{k_i}\theta_u^{k_u}\theta_c^{k_c}$, where θ_d and θ_u are the probability that a neighbor of u has not transmitted to u via in-degree edges and undirected edges respectively, given that it had not prior to time 0; θ_c is the probability that neither of the two neighbors of u in the same triangle has transmitted to u. Since we do not know k_u, k_i, k_o and k_c, then the probability that u is susceptible at time t can be written as

$$\begin{aligned} S(t) &= S(0)\sum_{k_i,k_o,k_u,k_c} p(k_i,k_o,k_u,k_c)\,\theta_d^{k_i}\theta_u^{k_u}\theta_c^{k_c} \\ &= S(0)\psi(\theta_d,1,\theta_u,\theta_c) \end{aligned} \tag{4.2}$$

Given $S(t)$, we are able to write equations for calculating E, I, and R based on the flow diagram in Figure 4.1.

$$\begin{aligned} E &= 1 - S - I - R \\ \dot{I} &= \sigma E - \gamma I \\ \dot{R} &= \gamma I \end{aligned} \tag{4.3}$$

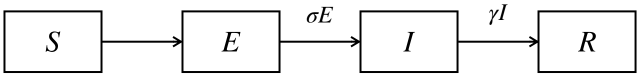

Figure 4.1: Flow diagram for the flux of individuals between different compartments. S, E, I, R represent the proportion of individuals in susceptible, exposed, infected, and recovered states. We have an explicit expression for S as shown in Equation (4.2), and the expressions for E, I, R are the same as in the *mass action* compartmental model.

Next, we need to calculate θ_u, θ_d and θ_c separately.

4.3.1.1 Considering $\boldsymbol{\theta_u}$

θ_u is the probability that a neighbor of u via an undirected edge has not transmitted infection to u at time t given that it has not at time 0. That is, $\theta_u(0) = 1$. It is divided into four parts (Figure 4.2): $\phi_{S,u}$, $\phi_{E,u}$, $\phi_{I,u}$, and $\phi_{R,u}$, which represent the probabilities that a neighbor of u is in susceptible, exposed, infected and recovered states respectively and has not transmitted infection to u. Thus,

$$\theta_u = \phi_{S,u} + \phi_{E,u} + \phi_{I,u} + \phi_{R,u} \tag{4.4}$$

$1 - \theta_u$ is the probability that a neighbor of u has transmitted infection to u. From the diagram in Figure 4.2, we have

$$\begin{aligned}
\dot{\theta}_u &= -\beta_u \phi_{I,u} \\
\dot{\phi}_{I,u} &= \sigma\phi_{E,u} - \gamma\phi_{I,u} - \beta_u\phi_{I,u} \\
\dot{\phi}_{E,u} &= -\dot{\phi}_{S,u} - \sigma\phi_{E,u}
\end{aligned} \tag{4.5}$$

Next, we need to find an expression for $\phi_{S,u}$ in terms of θs. Let's consider a neighbor of u with an undirected edge to u. The probability that the neighbor is susceptible at time 0 equals to the proportion of susceptible individuals in the population which is $S(0)$. The probability that the neighbor is attached to u by an undirected edge and has k_u undirected edges, k_i in-degree edges, k_o out-degree edges and k_c triangles is $k_u p(k_i, k_o, k_u, k_c)/\langle k_u \rangle$, where $\langle k_u \rangle = \sum_{k_i,k_o,k_u,k_c} k_u p(k_i, k_o, k_u, k_c)$, so the probability that an initially susceptible neighbor is still susceptible at time t is $S(0)k_u p(k_i, k_o, k_u, k_c)\theta_d^{k_i}\theta_u^{k_u-1}\theta_c^{k_c}/\langle k_u \rangle$ (u is prevented from infecting its neighbors, thus only $k_u - 1$ individual can infect the neighbor via undirected edges). Since we do not know k_u for the neighbor, the probability that a neighbor is susceptible at time t is written as

$$\begin{aligned} \phi_{S,u} &= \frac{\sum_{k_i,k_o,k_u,k_c} S(0)k_u p(k_i, k_o, k_u, k_c)\theta_d^{k_i}\theta_u^{k_u-1}\theta_c^{k_c}}{\sum_{k_u,k_i,k_o,k_c} k_u p(k_i, k_o, k_u, k_c)} \\ &= S(0)\,\frac{\frac{\partial}{\partial z}\psi(\theta_d, 1, \theta_u, \theta_c)}{\frac{\partial}{\partial z}\psi(1,1,1,1)} \end{aligned} \tag{4.6}$$

By combining Equations (4.5) and (4.6), we finish the system for θ_u.

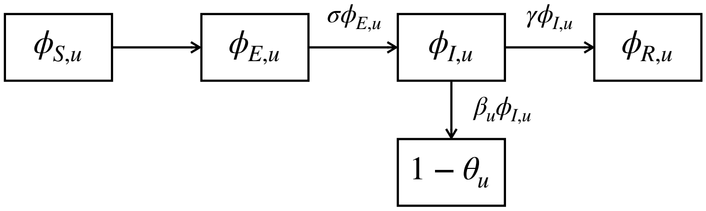

Figure 4.2: Flow diagram for the flux of neighbors of u connected by undirected edges through different states. $\phi_{S,u}$, $\phi_{E,u}$, $\phi_{I,u}$, $\phi_{R,u}$ and θ_u represent the probabilities that a neighbor of u connected by an undirected edge to u is susceptible, exposed, infected, recovered and has not transmitted infection to u. The sum of $\phi_{S,u}$, $\phi_{E,u}$, $\phi_{I,u}$ and $\phi_{R,u}$ equals to θ_u. $1-\theta_u$ is the probability that a neighbor of u has transmitted infection to u.

4.3.1.2 Considering $\boldsymbol{\theta_d}$

θ_d is the probability that a neighbor of u connected by directed edges has not transmitted infection to u at time t given that it had not at time 0. That is, $\theta_d(0)=1$. It is divided into four parts as shown in Figure 4.3. The derivation of expressions for θ_d, $\phi_{S,d}$, $\phi_{I,d}$, $\phi_{E,d}$ are similar to the infection transmission via undirected edges except that we need to consider the direction of directed edges explicitly. From the flux diagram in Figure 4.3, we have

$$\begin{aligned}
\theta_d &= \phi_{S,d}+\phi_{E,d}+\phi_{I,d}+\phi_{R,d}\\
\dot{\theta}_d &= -\beta_d\phi_{I,d}\\
\dot{\phi}_{I,d} &= \sigma\phi_{E,d}-\gamma\phi_{I,d}-\beta_d\phi_{I,d}\\
\dot{\phi}_{E,d} &= -\dot{\phi}_{S,d}-\sigma\phi_{E,d}
\end{aligned} \tag{4.7}$$

When deriving the equation for $\phi_{S,d}$, we consider a neighbor of u with a directed edge pointing to u. The probability that the neighbor is susceptible at time 0 is $S(0)$. The probability that the neighbor has k_i in-degree edges, k_o out-degree edges, k_u undirected edges and k_c triangles and reaches to u by an out-going edge is $k_o p(k_i, k_o, k_u, k_c)/\langle k_o \rangle$, where $\langle k_o \rangle = \sum_{k_i,k_o,k_u,k_c} k_o p(k_i, k_o, k_u, k_c)$. Given this, we have the probability that a neighbor of u is still susceptible at time t

$$\begin{aligned}\phi_{S,d} &= \frac{\sum_{k_i,k_o,k_u,k_c} S(0) k_o p(k_i, k_o, k_u, k_c) \theta_d^{k_i} \theta_u^{k_u} \theta_c^{k_c}}{\sum_{k_u,k_i,k_o,k_c} k_o p(k_i, k_o, k_u, k_c)} \\ &= S(0) \frac{\frac{\partial}{\partial y} \psi(\theta_d, 1, \theta_u, \theta_c)}{\frac{\partial}{\partial y} \psi(1,1,1,1)}\end{aligned} \tag{4.8}$$

By combining Equations (4.7) and (4.8), we finish the system for θ_d.

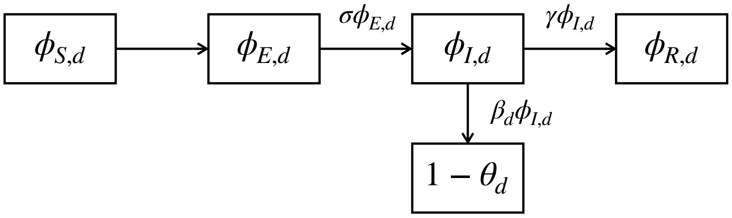

Figure 4.3: Flow diagram for the flux of neighbors of u connected by directed edges through different states. $\phi_{S,d}$, $\phi_{E,d}$, $\phi_{I,d}$, $\phi_{R,d}$, θ_d show probabilities that a neighbor of u through an directed edge is susceptible, exposed, infected, recovered and has not transmitted infection to u. The sum of $\phi_{S,d}$, $\phi_{E,d}$, $\phi_{I,d}$ and $\phi_{R,d}$ equals to θ_d. $1 - \theta_d$ is the probability that a neighbor of u has transmitted infection to u.

4.3.1.3 Considering $\boldsymbol{\theta_c}$

θ_c represents the probability that the two neighbors within the same triangle as u has not transmitted infection to u, given that they had not at time 0 ($\theta_c(0) = 1$). To calculate θ_c, we break it into 10 parts as shown in Figure 4.4. Each part donates to a combination of the states of the two neighbors within the same triangle. For example, ϕ_{SS} is defined as the probability that both neighbors are susceptible and has not transmitted infection to u; ϕ_{EI} indicates the probability that one neighbor is in exposed state and the other is in infected state, and neither has transmitted infection to u.

From the flux between different parts, we have

$$\begin{aligned}
\dot{\theta}_c &= -\beta_c\phi_{SI} - \beta_c\phi_{EI} - 2\beta_c\phi_{II} - \beta_c\phi_{IR} \\
\dot{\phi}_{SE} &= 2A\phi_{SS} - \sigma\phi_{SE} - A\phi_{SE} \\
\dot{\phi}_{SI} &= \sigma\phi_{SE} - \gamma\phi_{SI} - (\beta_c + A)\phi_{SI} - \beta_c\phi_{SI} \\
\dot{\phi}_{EE} &= A\phi_{SE} - 2\sigma\phi_{EE} \\
\dot{\phi}_{SR} &= \gamma\phi_{SI} - A\phi_{SR} \\
\dot{\phi}_{EI} &= (\beta_c + A)\phi_{SI} + 2\sigma\phi_{EE} - \beta_c\phi_{EI} - \sigma\phi_{EI} - \gamma\phi_{EI} \\
\dot{\phi}_{ER} &= A\phi_{SR} + \gamma\phi_{EI} - \sigma\phi_{ER} \\
\dot{\phi}_{II} &= \sigma\phi_{EI} - 2\gamma\phi_{II} - 2\beta_c\phi_{II} \\
\dot{\phi}_{IR} &= \sigma\phi_{ER} + 2\gamma\phi_{II} - \gamma\phi_{IR} - \beta_c\phi_{IR}
\end{aligned} \tag{4.9}$$

where A is the rate that a neighbor in a triangle is infected by individuals outside the triangle. To close the system for θ_c, we need to find expressions for A and ϕ_{SS}. When we consider each of the two neighbors in a triangle separately, the expression for the probability that one neighbor is still susceptible and has not transmitted infection to u is similar to that in the system for θ_u. Since we require both be susceptible, then

$$\phi_{SS} = \left(S(0) \frac{\frac{\partial}{\partial v}\psi(\theta_d, 1, \theta_u, \theta_c)}{\frac{\partial}{\partial v}\psi(1,1,1,1)} \right)^2 \tag{4.10}$$

From the diagram in Figure 4.4, we know $\dot{\phi}_{SS} = -2A\phi_{SS}$, which can be rewritten as

$$A = -\frac{\dot{\phi}_{SS}}{2\phi_{SS}} \tag{4.11}$$

By combination Equations from (4.2) to (4.11), we complete the network-based SEIR model.

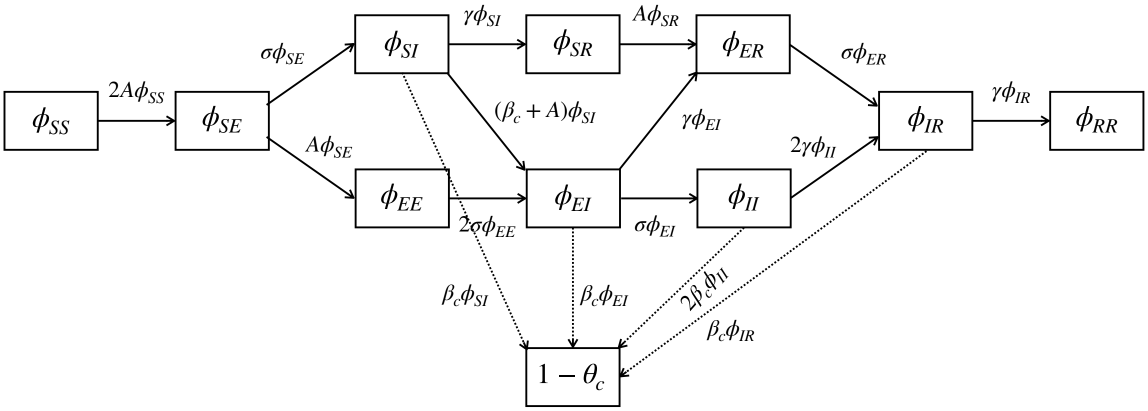

Figure 4.4: Flow diagram for the flux of two neighbors of u connected by the same triangle through different states. The flux shows the change between the probabilities that two neighbors of u within the same triangle are in different states, and have not transmitted infection to u. The sum of different states equals to θ_c. $1-\theta_c$ is the probability that one neighbor of u in the triangle has transmitted infection to u.

4.3.2 Model implementation and analysis

To explore the contribution of different population structures on infectious disease dynamics, we constructed a negative binomial distribution which allows us to keep the mean degree constant while change the variance of the distribution from the mean degree to infinity. The probability generating function (PGF) for a negative binomial distribution with parameters p and r is

$$\begin{aligned} g_{nb}(x;r,p) &= \sum_{k=0}^{\infty} P(k)x^{k+1} \\ &= \sum_{k=0}^{\infty} \binom{r+k-1}{k} p^r(1-p)^k x^{k+1} \end{aligned} \tag{4.12}$$

$$= \left(\frac{p}{1-(1-p)x}\right)^r x$$

Note the negative binomial distribution generated by Equation (4.12) starts from 1, instead of 0, to make sure no individual is isolated from the contact network. The distribution governs the distribution of edges from all population structures.

Following the approach described in [167], we modify the PGF so that all edges occur in pairs, and the degree will always be an even integer. That is, the number of pairs of edges follows a negative binomial distribution. We introduce two other parameters to assign each pair of edges to different population structures: with probability p_o, a pair of edges forms two out-degree edges with individuals that are not themselves connected; with probability p_c, a pair of edges is part of a triangle, and with probability $(1-p_o-p_c)$, a pair of edges forms two undirected edges with individuals who are not themselves connected. When $p_o = 0$, the network converges to a network without directed edges; when $p_c = 0$, the network has no clusters; when $p_o = p_c = 0$, the network only has undirected edges. We also assume that the proportion of HCWs in the population is α, and only HCWs have a constant number of in-degree edges. To keep the balance of in-degree edges and out-degree edges, we have

$$k_i = \frac{p_o k}{\alpha} \tag{4.13}$$

where k is the mean degree (pair of edges) of the network. Given these conditions, we have the PGF for the degree distribution of the network

$$\psi(x,y,z,v) = (\alpha x^{2k_i} + 1 - \alpha) \left(\frac{p}{1-(1-p)[p_o y^2 + (1-p_o-p_c)z^2 + p_c v]}\right)^r [p_o y^2 + (1-p_o-p_c)z^2 + p_c v] \tag{4.14}$$

By integrating Equations (4.14) with (4.2)--(4.11), we are able to model diseases transmission on a network with negative binomial distributed pairs of edges. In addition, we also generate a homogeneous network model where all individuals have equal number of edges and no triangles and directed edges exist in the network. The mean degree of the homogeneous network is equal to that of the negative binomial distributed network.

We assume there are 1,000 individuals in the network. The mean degree (k) of the network is 3, which is roughly equal to the average number of contacts per individual in Liberia estimated from contact tracing data during the 2014 Ebola epidemic [158]. We assume the proportion of out-degree edges per individual (p_o) is 0.2 if there are HCWs in the network; otherwise, $p_o = 0$. We consider p_c ranging from 0 to 1, and α ranging from 0.0 to 0.1.

In analyzing disease dynamics on the network, we fix the exposure rate $\sigma = 0.105$ and the recovery rate $\gamma = 0.122$, which roughly equal to the exposure rate and recovery rate of the Ebola virus disease estimated from the 2014 Ebola epidemic in Liberia [171]. We assume the transmission rates via undirected, directed and clustered edges are the same $(\beta_u = \beta_d = \beta_c = 0.2)$, and 1 randomly selected individual is infected at time 0.

We solved the model using a Real-valued Variable-coefficient Ordinary Differential Equation solver implemented in a Python library Scipy1.2.1 [172].

4.3.3 Data simulation and model fitting

To assess the ability of various models to infer transmission rates and make predictions based on data from an emerging outbreaks, we perform a simulation study in which we fit SEIR models with different combinations of population structures to simulated epidemiological data. Simulated incidence data is generated from the network-based SEIR model with three population structures (heterogeneous contact pattern, clustering, and directed contact) and Poisson distributed noise. We assume the degree distribution of the network follows a negative binomial distribution as described in 4.3.2. The distribution has a mean (pairs of edges) of 3, which roughly equals to the number of contacts per individual in Libera, and a variance of 4. The distribution is approximate to an exponential distribution which is common in empirical contact networks [164]. The network contains 10,000 individuals, and has a cluster coefficient of 0.7, which is taken from [158]. That is, the proportion of edges being part of triangles is 0.7. The proportion of HCWs in the population is 0.00048, which is close to the proportion of HCWs in Liberia [173], and the proportion of out-degree edges per individual is 0.2. Initially, 10 randomly selected individual are infected. The exposure rate σ and recovery rate γ equal to 0.105 and 0.122, respectively. The transmission rate through undirected edges (β_u) is 0.1, and the transmission rate via directed edges and triangles (β_a, β_c) is 0.5.

We fit five different models to the first n-days cumulative incidence of simulated epidemic, where $n = 20, 30$: (1) a heterogeneous network model including undirected, directed and clustered edges; (2) a heterogeneous network model only having undirected and directed edges; (3) a heterogeneous network model only including undirected and clustered edges; (4) a heterogeneous network model only having undirected edges; (5) a homogeneous network model in which each individual has the same number of contacts.

We first fit each model to available data to estimate the transmission rate, and assume the transmission rates via different edges are consistent. All other model parameters are fixed at their true values, unless certain parameters do not exist in corresponding models. Parameter estimation is accomplished by minimizing the sum of squared errors between simulated and predicted cumulative incidence using the *minimize* function in the Python library *lmfit* [174]. Then we use each model with estimated transmission rate to project epidemic dynamic. We assess the performance of each model in terms of three different surveillance targets: (1) Relative total incidence (the ratio of predicted total incidence over ground truth); (2) Relative peak intensity (the ratio of predicted peak intensity over ground truth); (3) Relative timing of the epidemic peak (the lag between the predicted timing and ground truth).

4.4 RESULTS

4.4.1 The contribution of different population structures on infectious disease dynamics.

In general, heterogeneity in contact pattern (controlled by variance of a degree distribution) decreases the total incidence, and the time reaching the peak of an epidemic (Figure 4.5(A)). However, the heterogeneity of a network has little effect on the peak intensity. When a proportion of undirected edges is turned into clusters (triangles), there is minor effect on total incidence, while the peak intensity of an epidemic decreases and the peaking timing becomes later (Figure 4.5(B)). If HCWs are included in a network and each individual has out-degree edges directed to HCWs, it decreases the total incidence and peak intensity, while does not affect the peak timing (Figure 4.5(C)).

A systematic evaluation of the three population structures (heterogeneous contact pattern, directed contacts, and clustering) shows that heterogeneous contact pattern reduces the total incidence and epidemic peak timing monotonically (Figures 4.6(A)(C), 4.S1, 4.S3). A larger variance leads to a smaller total incidence, and a short time reaching the peak of an epidemic further. When comparing to a homogeneous model (without any population structures), the total incidence generated by an undirected network model (with heterogeneous contacts) is always smaller while the peak is always reached earlier (Figure 4.6(A)(C)). The divergence becomes larger with the variance increasing. In contrast, the effect of heterogeneity on peak intensity is complex, which depends on the variance of the degree distribution. The peak intensity increases when the variance ranges from 2 to 6, and starts to decrease after the variance is over 6 (Figures 4.6(B) and 4.9). The peak intensity generated by an undirected network model is slightly higher than that generated by a homogeneous model when the variance is small, while it becomes lower when the variance is around 22 (approximate to a scale-free network). However, the difference of peak intensity between an undirected network model and a homogeneous model is not significant.

When directed contacts are integrated into the network model, it is able to decrease the total incidence further (Figure 4.6(A)). It might stem from the direction of those directed edges, as directed contacts only point to HCWs that makes HCWs become a sink of the disease. However, when changing the proportion of HCWs in a population from 0.00048 to 0.1, it does not change the total incidence significantly (Figure 4.8). It suggests that a better prediction might be achieved, even if the true proportion of HCWs in a certain location is unknown, by using a value of α from other locations instead of ignoring it. Meanwhile, directed edges also lower the peak intensity significantly (Figure 4.6(B)). The peak intensity keeps increasing continuously if there are more HCWs in a population (Figure 4.9). Unlike the effect of heterogeneous contacts on peak timing, directed edges deaccelerate the speed reaching to the epidemic peak (Figure 4.6(C)).

When we explore the effect of clustering on total incidence, we found that it does decrease the total incidence slightly when clustered contacts are included in a network model (Figure 4.6(A)). However, the effect does not change significantly as the clustering increases (Figure 4.8), especially when the variance ranges from 2 to 5. This is consistent with a previous study [167]. In terms of peak intensity and timing, the effect of clustering is similar to that of directed edges – it decreases the peak intensity and timing when being introduced into a network model (Figure 4.6(B)(C)). However, the peak intensity decreases with clustering increasing, which is opposite to the effect of directed contacts (Figure 4.9).

We also examine the contribution of the three population structures on total incidence, peak intensity and timing. We quantify the contributions of heterogeneous, directed and clustered edges by comparing the difference of prediction between the homogeneous model and the undirected network mode, the undirected network model and semi-directed network model, the semi-directed network model and the full model (with all three population structures), respectively. We found that total incidence is mainly driven by heterogeneity and directed contacts (Figure 4.6(A)). When variance of the degree distribution is less than 10, the contribution of directed contacts is over that of heterogeneous contact pattern; otherwise, they are similar to each other. In terms of peak intensity, heterogeneous contact pattern has minimum contribution, while directed contacts has the largest contribution (Figure 4.6(B)). The contribution of clustering to both total incidence and peak intensity is minor. In contrast, heterogeneous contact pattern dominates and contributes positively to peak time (accelerate the speed reaching epidemic peak), whereas directed and clustered contacts have a negative contribution to the peak time (deaccelerate the pace reaching epidemic peak slightly). The positive contribution of heterogeneous contact pattern is much larger than the negative contribution of both directed and clustered edges. Overall, the result suggests that in modeling an epidemic, we should at least include heterogeneous and directed contacts in a network-based SEIR model.

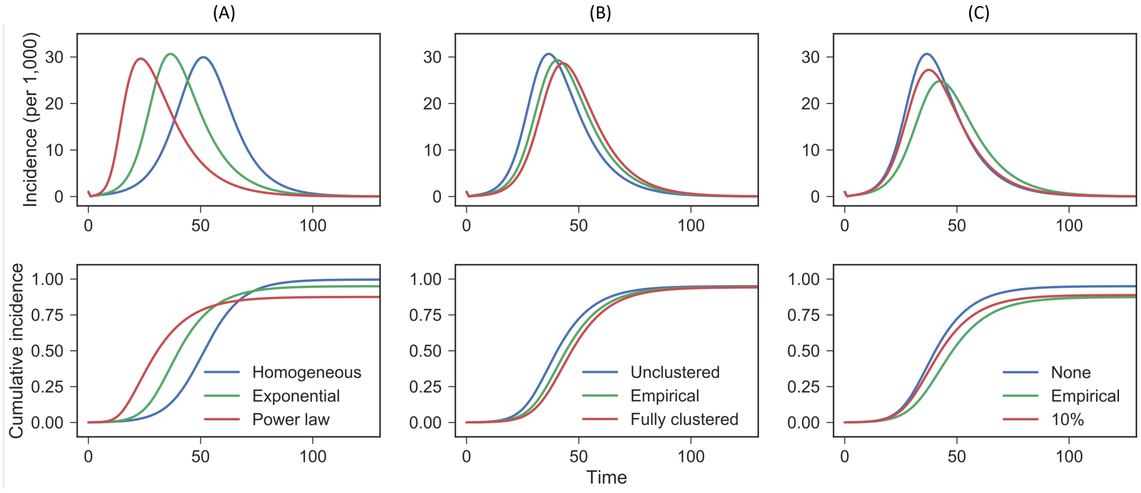

Figure 4.5: Epidemic curves generated by SEIR models with different network structures. The mean degree (pairs of edges) of each network is 3. (A) The effect of heterogeneous contact pattern on epidemic dynamics. 'Homogeneous' represents that each individual in a network has equal number of contacts; 'Power law' represents a scale-free network, which has very large variance; 'Exponential' means the degree distribution of a network follows an exponential distribution and its variance is between a homogeneous network and a scale-free one. (B) The effect of clustering on epidemic dynamics. 'Unclustered' represents a network without clustered edges; 'Empirical' indicates a network with 70% of all edges being part of clusters (triangles) which is roughly equal to the one estimated from Ebola epidemic [158]; 'Full clustered' is a network in which all edges are clustered. (C) The effect of directed contacts on epidemic dynamics. 'None' indicates there is no HCWs in a network and thus has no directed edges; 'Empirical' is a network that the proportion of HCWs in the population is 0.00048, which is equivalent to the ratio of HCWs to the total population in Liberia; '10%' means that 10% of the individuals is HCWs.

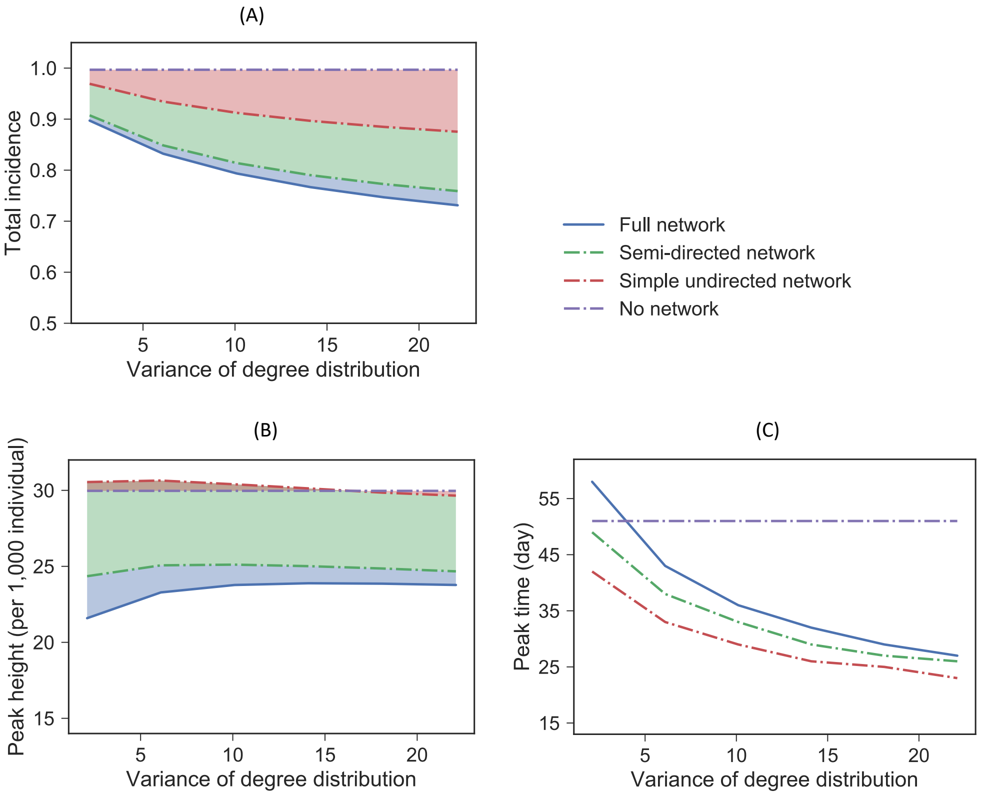

Figure 4.6: Contributions of different network structures to (A) total incidence, (B) peak intensity, and (C) peak timing. ‘Full network’ is a network model including all three population structures (heterogeneous, clustered and directed contacts); ‘semi-directed network’ is a network model including heterogeneous and directed contacts; ‘Simple undirected network’ is a network model including only heterogeneous contacts; ‘No network’ is a homogeneous network model in which each individual has the equal number of contacts; ‘Clustered undirected network’ is a network model including both clustered and heterogeneous contacts. Pairs of edges in each network follow a negative binomial distribution with a mean degree of 3 and a variance ranging from 2.1 to 22.1. The ratios of clustered edges to all edges per individual is 0.7 if exists in a network model; the proportion of HCWs in the population is 0.00048 if directed contacts exists in a network model.

4.4.2 Epidemic prediction using network models with various population structures

When an epidemic occurs, public health agencies usually plan resource allocation, hospital bed capacity and the distribution of antiviral drugs in the initial phase of an epidemic based on model predicted epidemic trajectory. However, we usually do not know actual values of transmission parameters, which have to be estimated from data collected in an epidemic prior to making predictions.

To investigate how complex a SEIR model should be to make accurate predictions for an epidemic, we fit five different models with various population structures to simulated data as described in Section 4.3.3 and then make future prediction for the epidemic. We consider a scenario where the transmission rate via undirected edges is 0.1, while it is 0.5 via directed and clustered edges. This assumption appropriates for diseases transmitted via directed human-to-human contact via body fluids or bloods, such as Ebola, in which transmission rates within hospital and household are usually higher.

We find that the total incidence and peak intensity predicted by a homogeneous model (without population structures), which is the most popular model used in practice, always has the largest bias no matter how much data is available (Figure 4.7, Table 4.2). The largest differences of the total incidence and peak intensity between prediction and ground truth reach 11% and 22%, respectively. The homogeneous model also predict the peak timing 6 days earlier when only data prior to the exponential growth phase is available, while the prediction on peak timing is close to ground truth when it is made at the end of the exponential growth phase (Figure 4.7, Table 4.2). Similarly, when making predictions using an undirected network model (with heterogeneity), it predicts 6% more cases for the epidemic (Table 4.2). In terms of peak intensity, the prediction made by a clustered network model (with heterogeneous and clustered contacts) is slightly higher than ground truth regardless of the data availability. Otherwise, the prediction on peak intensity and timing by the undirected network model and the clustered network model do not diverge from ground truth significantly (Figure 4.7, Table 4.2). In contrast, a full network model (with all three population structures) and a semi-directed network (with undirected and directed contacts) model make the most accurate prediction on total incidence, peak intensity and timing, except that the predicted peak intensity by the semi-directed network model is 8% lower than ground truth when 20-days incidence data is available. Our results suggest that for an epidemic in real world, a homogeneous model should not be used for predicting the epidemic trajectory, and the effects of population structures cannot be cancelled completely by adjusting transmission rates during the parameter estimation process. We should at least incorporate heterogeneous contacts in an epidemiological model for making reliable predictions.

Table 4.2: Performance of SEIR models with different population structures on predicting the total incidence, peak intensity and timing

Model	20-days data available				30-days data available			
	β_{est}	S_{rel}	H_{rel}	T_{rel}	β_{est}	S_{rel}	H_{rel}	T_{rel}
Homogeneous network	0.298	1.11	1.22	6	0.343	1.11	1.29	3
Undirected network	0.188	1.05	1.00	3	0.2	1.06	1.04	2
Clustered & undirected network	0.247	1.07	1.09	1	0.249	1.07	1.09	1
Semi-directed network	0.243	0.99	0.92	3	0.269	1.00	0.97	1
Full network	0.362	1.00	0.99	2	0.383	1.00	1.01	1

Note: β_{est} donates to the estimated per contact transmission rate through edges; S_{rel}, H_{rel}, T_{rel} represent the relative total incidence, peak intensity and peak timing of the corresponding model to ground truth, respectively.

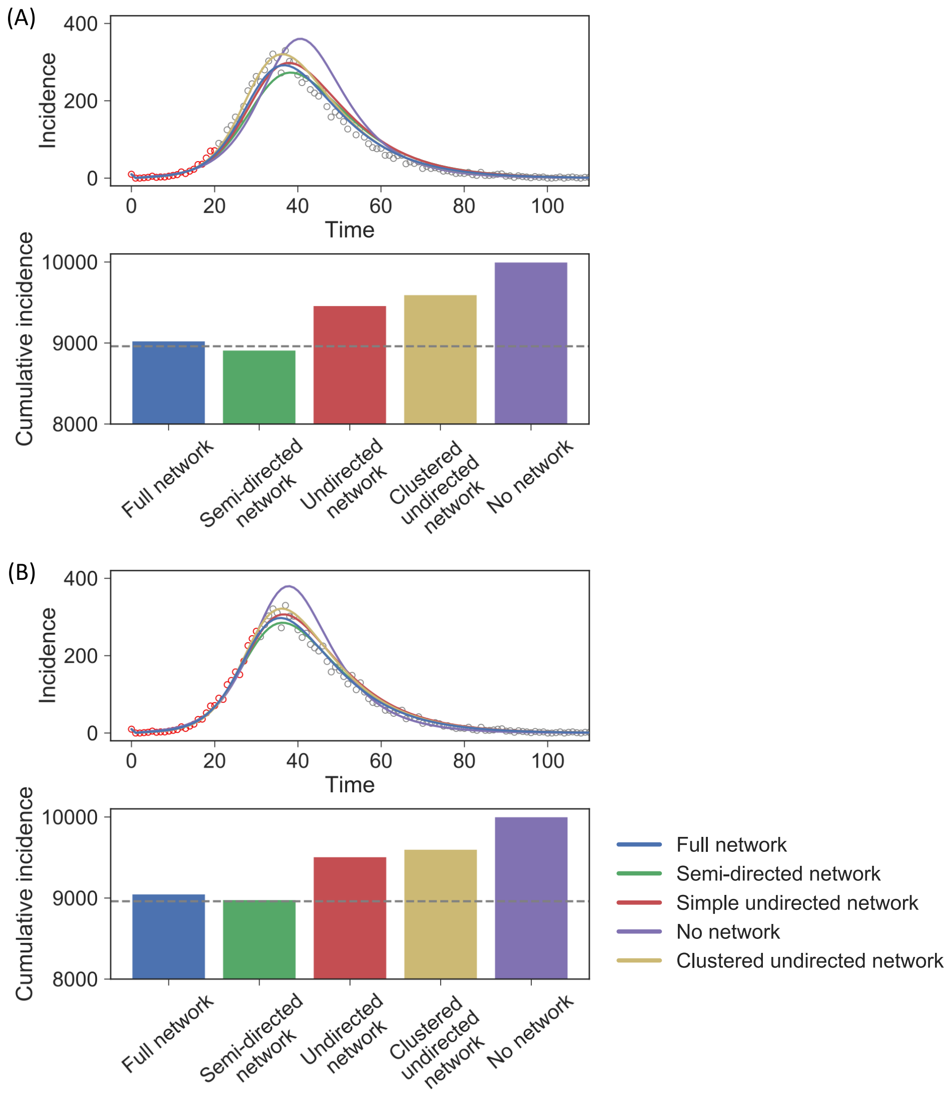

Figure 4.7: The effect of network structures on epidemic predictions. The per contact transmission rate via undirected edges is 0.1; both transmission rates through directed and clustered edges equal to 0.5. Circles indicate simulated data of an epidemic using the network-based SEIR model (Equations (4.2)--(4.11) and (4.13)--(4.14)) with Poisson distributed noise. Each network model is fitted to data from day 0 to (A) day 20 and (B) day 30 (circles in red), respectively to estimate the transmission rate. We assume the transmission rates through different edges are equal in model fitting process. Each curve represents the predicted epidemic by the corresponding model with the estimated value of the transmission rate. Each bar shows the predicted total incidence by the corresponding mode. Grey horizontal line is the true total incidence of simulated epidemic.

4.5 DISCUSSION

In the study, we have derived a network-based SEIR model using an edge-based compartmental modeling approach. The model includes three different population structures – heterogeneous contact pattern, directed contacts and clustering, in which transmission rates can be various through different edges. Using this model, we investigated the contribution of different structures on infectious disease dynamics. In an ideal scenario where the contact network and parameters in an epidemic are known in advance, we find that total incidence, peak intensity and timing are driven by different factors. For example, the results suggest that total incidence is dominated by both heterogeneous and directed contacts. Without these two structures, the model overestimates the total incidence significantly, and the contribution of heterogeneous contact pattern keep increasing as variance of the degree distribution increases. In contrast, directed contacts is the only main factor impacting the peak intensity of an epidemic. In terms of the peak timing, heterogeneous contact pattern accelerates the speed of reaching the epidemic peak, while the other two structures slow down the pace slightly. Even though clustering has no significant effect on dynamics, it does reduce epidemic risk.

However, parameters relevance to infectious diseases, especially transmission rates, are not available in real world. To predict epidemic dynamics, a common practice is to first estimate those parameters from epidemiological data using mathematical models and then make predictions. A challenging problem is what network structures should be incorporated in a model. There is always a tradeoff between model complexity and model accuracy. Theoretically, a more complex model should make more accurate predictions, whereas it possesses more parameters, including those defining population structures. It takes extra time to obtain values of the parameters from other sources or estimate them with transmission parameters simultaneously, making the model fitting challenging. In the study, we explore how accurate a prediction could be by models in different levels of complexity. We consider a simple scenario where we have already known everything about network structures and only need to estimate transmission rates. Our results show that a homogeneous model always overestimates total incidence, peak intensity and timing significantly. This indicates that the overestimation of the total incidence in 2014 Ebola epidemic might stem from not only the effective public health interventions but also the inappropriate model usage in prediction. Our study suggests that heterogeneous contact patterns should be at least included in a model to make reliable predictions.

The study provides insights for public health agencies about contributions of different population structures on epidemics, and also proves that predictions made by a homogeneous is not reliable in practice. The extension of a SEIR model to a network with multiple structures will allow us to build and rapidly analyze infectious disease transmission on more realistic models. The model can also be used to analyze the effectiveness of different intervention strategies, and where these strategies should be implemented in an emerging epidemic to maximize the effect of interventions. In the future, population structures in other dimensions can be introduced on top of this model, such as heterogeneity between and within different age groups, and serosorting [169] *etc*..

This study has a couple limitations. First, our model assumes all clusters in a population are triangles and they have no shared edges. However, clusters are more than triangles in real world. For example, in a household with 5 individuals, interactions might occur between any two of them. Even though triangles are the smallest *clique* describing the cluster, there could be shared edges between any two triangles which decreases the total number of edges within a network. To relax the assumption, we can follow a motif-based generalization of the configuration model that allows triangles and other cliques to share edges [167,175]. Second, when fitting models to epidemiological data, we assume that parameters relevance to network structures are already known, which is hardly possible in practice. To overcome this drawback, we suggest public health agencies to collect and make contract tracing data freely available to epidemiologists as early as possible during an epidemic, so that epidemiologists are able to estimate network parameters and apply them to model predictions. In case there is no contact tracing data available, we suggest to use exponential distribution as the degree distribution in a population, as a previous study showed [164], and estimate the single parameter of the distribution from epidemiological data. Since parameters in a degree distribution might be correlated with the transmission rates, the approach described in [158] could be a good option for estimating correlated parameters from epidemiological data.

4.6 Supplemental information

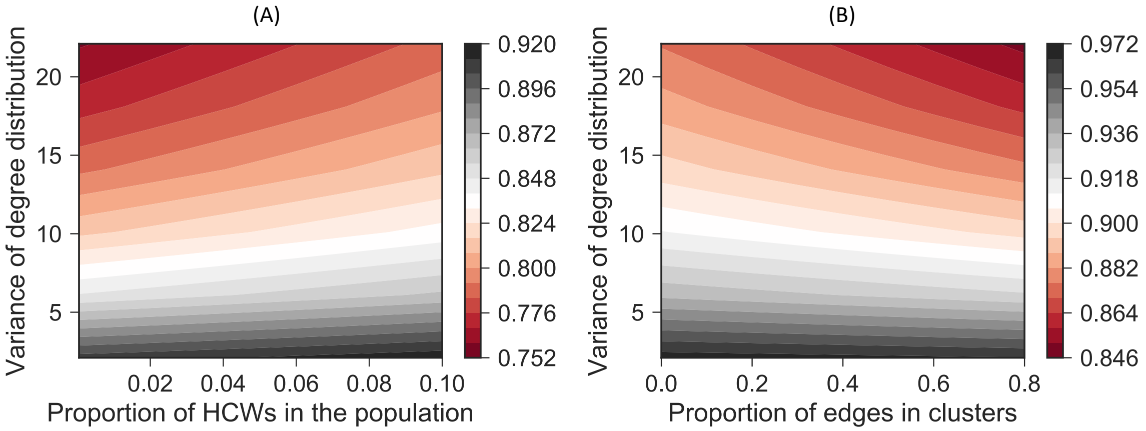

Figure 4.8: Joint effects of heterogeneous contact pattern (variance of degree distribution) and (A) directed contacts (proportion of HCWs in a population) (B) clustering (proportion of edges being part of triangles), respectively, on total incidence. Transmission rates via undirected, directed and clustered edges equal to 0.1. Both results are generated using the network-based SEIR model (Equations (4.2)--(4.11) and (4.13)--(4.14)). (A) Parameters in the model: $\alpha = 0.00048$, $p_o = 0.2$, $p_c = 0$. (B) Parameters in the model: $\alpha = 0$, $p_o = 0$, $p_c = 0.7$.

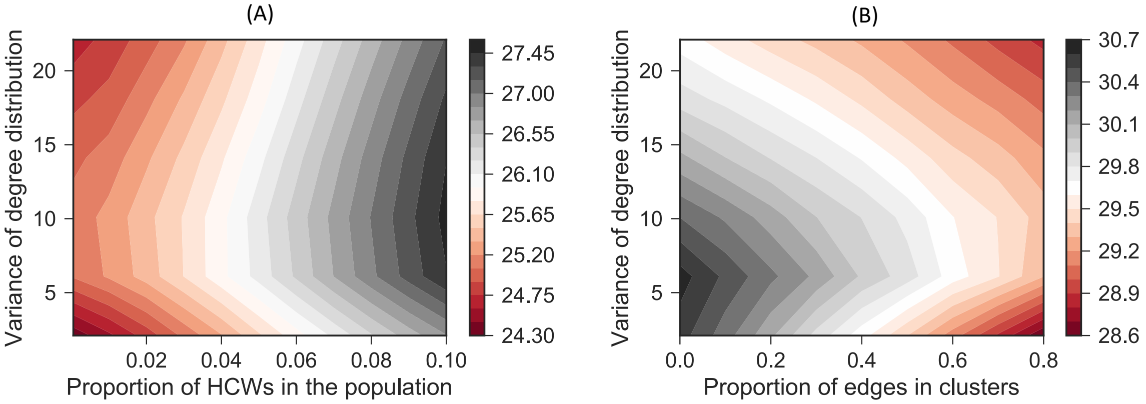

Figure 4.9: Joint effects of heterogeneous contact pattern (variance of degree distribution) and (A) directed contacts (proportion of HCWs in a population) (B) clustering (proportion of edges being part of triangles), respectively, on peak intensity of an epidemic. Transmission rates via undirected, directed and clustered edges equal to 0.1. Both results are generated using the network-based SEIR model (Equations (4.2)--(4.11) and (4.13)--(4.14)). (A) Parameters in the model: $\alpha = 0.00048$, $p_o = 0.2$, $p_c = 0$. (B) Parameters in the mode: $\alpha = 0$, $p_o = 0$, $p_c = 0.7$.

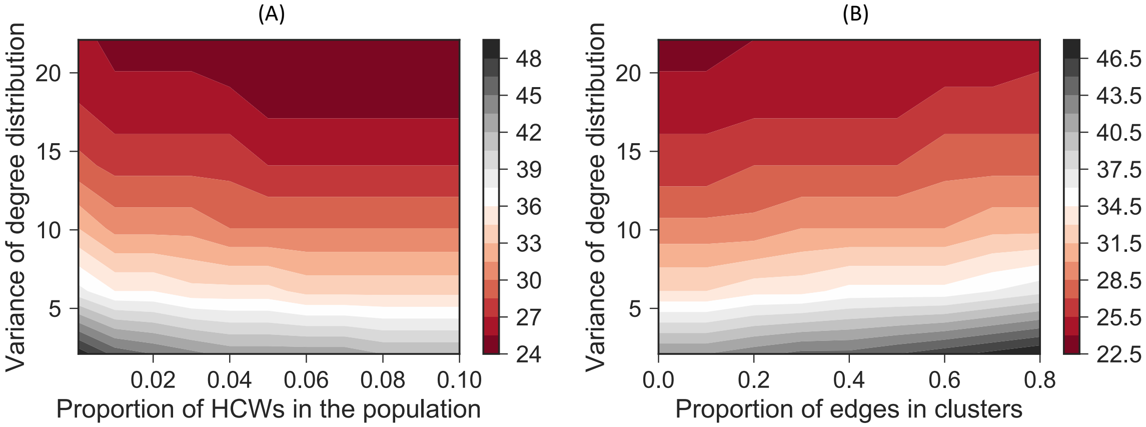

Figure 4.10: Joint effects of heterogeneous contact pattern (variance of degree distribution) and (A) directed contacts (proportion of HCWs in the population) (B) clustering (proportion of edges being part of triangles), respectively, on peak timing of an epidemic. Transmission rates via undirected, directed and clustered edges equal to 0.1. Both results are generated using the network-based SEIR model (Equations (4.2)--(4.11) and (4.13)--(4.14)). (A) Parameters in the model: $\alpha = 0.00048$, $p_o = 0.2$, $p_c = 0$. (B) Parameters in the model: $\alpha = 0$, $p_o = 0$, $p_c = 0.7$.

Bibliography

1. Biggerstaff M, Alper D, Dredze M, Fox S, Fung IC-H, Hickmann KS, et al. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. BMC Infect Dis. BioMed Central; 2016;16: 357. doi:10.1186/s12879-016-1669-x

2. Biggerstaff M, Johansson M, Alper D, Brooks LC, Chakraborty P, Farrow DC, et al. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. Epidemics. Elsevier; 2018;24: 26–33. doi:10.1016/J.EPIDEM.2018.02.003

3. Center for Disease Control and Prevention. Epidemic Prediction Initiative [Internet]. [cited 4 Jun 2019]. Available: https://predict.cdc.gov/

4. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. Proc Natl Acad Sci U S A. 2012;109: 20425–30. doi:10.1073/pnas.1208772109

5. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012-2013 season. Nat Commun. Nature Publishing Group; 2013;4: 2837. doi:10.1038/ncomms3837

6. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible Modeling of Epidemics with an Empirical Bayes Framework. PLoS Comput Biol. 2015;11: 1–18. doi:10.1371/journal.pcbi.1004382

7. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013–2014 Influenza Season Using Wikipedia. Salathé M,

editor. PLOS Comput Biol. Springer; 2015;11: e1004239. doi:10.1371/journal.pcbi.1004239

8. México Dirección General Adjunta de Epidemiología. Brote de influenza humana A H1N1 México. 2009.

9. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. Science. American Association for the Advancement of Science; 2009;324: 1557–61. doi:10.1126/science.1176062

10. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, et al. Emergence of Zaire Ebola Virus Disease in Guinea. N Engl J Med. Massachusetts Medical Society; 2014;371: 1418–1425. doi:10.1056/NEJMoa1404505

11. Olympus High Performance Compute Cluster at Carnegie Mellon University, supported by National Institute of General Medical Sciences Modeling Infectious Disease Agent Study (MIDAS) Informatics Services Group grant 1U24GM110707 [Internet]. 2018. Available: https://www.psc.edu/resources/computing/olympus

12. Zhang Q, Sun K, Chinazzi M, y Piontti AP, Dean NE, Rojas DP, et al. Spread of Zika virus in the Americas. Proc Natl Acad Sci. National Acad Sciences; 2017;114: E4334--E4343.

13. Shewhart WA. Economic control of quality of manufactured product. ASQ Quality Press; 1931.

14. Page ES. Continuous inspection schemes. Biometrika. JSTOR; 1954;41: 100–115.

15. Lorden G, others. Procedures for reacting to a change in distribution. Ann Math Stat. Institute of Mathematical Statistics; 1971;42: 1897–1908.

16. Roberts SW. Control chart tests based on geometric moving averages. Technometrics. Taylor & Francis Group; 1959;1: 239–250.

17. Fricker RD. Introduction to statistical methods for biosurveillance: with an emphasis on syndromic surveillance. Cambridge University Press; 2013.

18. Boyle JR, Sparks RS, Keijzers GB, Crilly JL, Lind JF, Ryan LM. Prediction and surveillance of influenza epidemics. Med J Aust. Wiley Online Library; 2011;194: S28--S33.

19. Mathes RW, Lall R, Levin-Rector A, Sell J, Paladini M, Konty KJ, et al. Evaluating and implementing temporal, spatial, and spatio-temporal methods for outbreak detection in a local syndromic surveillance system. Codeço CT, editor. PLoS One. Public Library of Science; 2017;12: e0184419. doi:10.1371/journal.pone.0184419

20. Pervaiz F, Pervaiz M, Rehman NA, Saif U. FluBreaks: early epidemic detection from Google flu trends. J Med Internet Res. JMIR Publications Inc.; 2012;14.

21. Cowling B, Wong I, Ho L, Riley S, Leung G. Methods for monitoring influenza surveillance data. Int J Epidemiol. 2006; Available: http://ije.oxfordjournals.org/content/35/5/1314.short

22. Griffin BA, Jain AK, Davies-Cole J, Glymph C, Lum G, Washington SC, et al. Early detection of influenza outbreaks using the DC Department of Health's syndromic surveillance system. BMC Public Health. BioMed Central; 2009;9: 483. doi:10.1186/1471-2458-9-483

23. Pelecanos AM, Ryan PA, Gatton ML. Outbreak detection algorithms for seasonal disease data: a case study using ross river virus disease. BMC Med Inform Decis Mak. BioMed Central; 2010;10: 74.

24. Watkins RE, Eagleson S, Veenendaal B, Wright G, Plant AJ. Applying cusum-based methods for the detection of outbreaks of Ross River virus disease in Western Australia. BMC Med Inform Decis Mak. BioMed Central; 2008;8: 37. doi:10.1186/1472-6947-8-37

25. Li Z, Lai S, Buckeridge DL, Zhang H, Lan Y, Yang W. Adjusting outbreak detection algorithms for surveillance during epidemic and non-epidemic periods. J Am Med Informatics Assoc. BMJ Group BMA House, Tavistock Square, London, WC1H 9JR; 2011;19: e51--e53.

26. Zhang H, Lai S, Wang L, Zhao D, Zhou D, Lan Y, et al. Improving the performance of outbreak detection algorithms by classifying the levels of disease incidence. PLoS One. Public Library of Science; 2013;8: e71803.

27. Lai S, Li X, Zhang H. Early Detection for Hand, Foot, and Mouth Disease Outbreaks. Early Warning for Infectious Disease Outbreak. Elsevier; 2017. pp. 283–294. doi:10.1016/B978-0-12-812343-0.00015-1

28. Wieland SC, Brownstein JS, Berger B, Mandl KD. Automated real time constant-specificity surveillance for disease outbreaks. BMC Med Inform Decis Mak. BioMed Central; 2007;7: 15. doi:10.1186/1472-6947-7-15

29. Spanos A, Theocharis G, Karageorgopoulos DE, Peppas G, Fouskakis D, Falagas ME. Surveillance of Community Outbreaks of Respiratory Tract Infections Based on House-Call Visits in the Metropolitan Area of Athens, Greece. Drews SJ, editor. PLoS One. Public Library of Science; 2012;7: e40310. doi:10.1371/journal.pone.0040310

30. Karami M, Ghalandari M, Poorolajal J, Faradmal J. Early Detection of Meningitis Outbreaks: Application of Limited-baseline Data. Iran J Public Health. Tehran University of Medical Sciences; 2017;46: 1366–1373. Available: http://www.ncbi.nlm.nih.gov/pubmed/29308380

31. Kammerer JS, Shang N, Althomsons SP, Haddad MB, Grant J, Navin TR. Using statistical methods and genotyping to detect tuberculosis outbreaks. Int J Health Geogr. BioMed Central; 2013;12: 15. doi:10.1186/1476-072X-12-15

32. Pervaiz F, Pervaiz M, Abdur Rehman N, Saif U. FluBreaks: early epidemic detection from Google flu trends. J Med Internet Res. JMIR Publications Inc.; 2012;14: e125. doi:10.2196/jmir.2102

33. Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response early aberration reporting system (EARS). J Urban Heal. Springer; 2003;80: i89--i96.

34. Bradley CA, Rolka H, Walker D, Loonsk J. BioSense: implementation of a national early event detection and situational awareness system. MMWR Morb Mortal Wkly Rep. 2005;54: 11–19.

35. Bravata DM, McDonald KM, Smith WM, Rydzak C, Szeto H, Buckeridge DL, et al. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. Ann Intern Med. Am Coll Physicians; 2004;140: 910–922.

36. Shmueli G, Burkom H. Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. doi:10.1198/TECH.2010.06134

37. Santillana M, Nguyen AT, Louie T, Zink A, Gray J, Sung I, et al. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. Sci Rep. Nature Publishing Group; 2016;6: 25732. doi:10.1038/srep25732

38. Chunara R, Aman S, Smolinski M, Brownstein JS. Flu near you: an online self-reported influenza surveillance system in the USA. Online J Public Health Inform. 2013;5.

39. Brownstein JS, Chu S, Marathe A, Marathe M V, Nguyen AT, Paolotti D, et al. Combining Participatory Influenza Surveillance with Modeling and Forecasting: Three Alternative Approaches. JMIR public Heal Surveill. JMIR Publications Inc.; 2017;3: e83. doi:10.2196/publichealth.7344

40. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. Nature Publishing Group; 2009;457: 1012–4. doi:10.1038/nature07634

41. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol. Public Library of Science; 2014;10: e1003581.

42. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One. Public Library of Science; 2013;8: e83672.

43. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. Am J Trop Med Hyg. ASTMH; 2012;86: 39–45.

44. Chan EH, Sahai V, Conrad C, Brownstein JS, Beatty M, Stone A, et al. Using Web Search Query Data to Monitor Dengue Epidemics: A New Model for Neglected Tropical Disease Surveillance. Aksoy S, editor. PLoS Negl Trop Dis. Public Library of Science; 2011;5: e1206. doi:10.1371/journal.pntd.0001206

45. Joner Jr MD, Woodall WH, Reynolds Jr MR, Fricker Jr RD. A one-sided MEWMA chart for health surveillance. Qual Reliab Eng Int. Wiley Online Library; 2008;24: 503–518.

46. Allen MB, Isaacson EL. Numerical analysis for applied science. John wiley & sons; 2011.

47. Centers for Disease Control and Prevention. Fluview [Internet]. 2018. Available: https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

48. Google. Data Source: Google Correlate [Internet]. [cited 12 Oct 2017]. Available: http://www.google.com/trends/correlate

49. Google. Data Source: Google Trends [Internet]. 2018 [cited 12 Oct 2017]. Available: https://trends.google.com/trends/

50. Farrow D. Modeling the past, present, and future of influenza. Private Communication, Farrow's work will be part of the published Phd thesis at Carnegie Mellon University. 2016.

51. Centers for Disease Control and Prevention. Overview of influenza surveillance in the United States. In: Fact Sheet [Internet]. 2006. Available: https://www.cdc.gov/flu/weekly/overview.htm

52. Centers for Disease Control and Prevention. 2009-2010 Influenza (Flu) Season [Internet]. 2019. Available: https://www.cdc.gov/flu/pastseasons/0910season.htm

53. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. Nature Publishing Group; 2009;457: 1012.

54. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012--2013 season. Nat Commun. Nature Publishing Group; 2013;4: 2837.

55. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible modeling of epidemics with an empirical Bayes framework. PLoS Comput Biol. Public Library

of Science; 2015;11: e1004382.

56. Ertem Z, Raymond D, Meyers LA. Optimal multi-source forecasting of seasonal influenza. PLoS Comput Biol. Public Library of Science; 2018;14: e1006236.

57. Dasey T, Reynolds HD, Nurthen N, Kiley C, Silva J. Biosurveillance ecosystem (bsve) workflow analysis. Online J Public Health Inform. University of Illinois at Chicago Library; 2013;5.

58. World Health Organization (WHO). Dengue and severe dengue [Internet]. [cited 26 May 2019]. Available: https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue

59. Stanaway JD, Shepard DS, Undurraga EA, Halasa YA, Coffeng LE, Brady OJ, et al. The global burden of dengue: an analysis from the Global Burden of Disease Study 2013. Lancet Infect Dis. Elsevier; 2016;16: 712–723. doi:10.1016/S1473-3099(16)00026-8

60. Center for Disease Control and Prevention (CDC). Dengue in Puerto Rico [Internet]. [cited 19 Feb 2019]. Available: https://www.cdc.gov/dengue/about/inpuerto.html

61. Pan American Health Organization (PAHO). Number of reported Dengue fever infection cases in Mexico from 2015 to 2018 [Internet]. [cited 19 Feb 2019]. Available: https://www.statista.com/statistics/939846/reported-dengue-fever-infection-cases-mexico/

62. PERÚ Instituto Nacional de Estadística e Informática (INEI). Number of reported cases of Dengue fever infection in Peru from 2007 to 2019 [Internet]. [cited 19 Feb 2019]. Available: https://www.statista.com/statistics/821068/number-reported-cases-dengue-fever-peru/

63. Gubler DJ. How Effectively is Epidemiological Surveillance Used for Dengue Programme Planning and Epidemic Response? [Internet]. Dengue Bulletin. 2002. Available: https://apps.who.int/iris/bitstream/handle/10665/163775/dbv26p96.pdf;jsessionid=F13B73F8500E17EC62C4DD7AC044DE76?sequence=1

64. Beatty ME, Stone A, Fitzsimons DW, Hanna JN, Lam SK, Vong S, et al. Best practices in dengue surveillance: a report from the Asia-Pacific and Americas Dengue Prevention Boards. PLoS Negl Trop Dis. Public Library of Science; 2010;4: e890. doi:10.1371/journal.pntd.0000890

65. Vong S, Goyet S, Ly S, Ngan C, Huy R, Duong V, et al. Under-recognition and reporting of dengue in Cambodia: a capture–recapture analysis of the National Dengue Surveillance System. Epidemiol Infect. Cambridge University Press; 2012;140: 491–499. doi:10.1017/S0950268811001191

66. Wichmann O, Yoon I-K, Vong S, Limkittikul K, Gibbons R V, Mammen MP, et al. Dengue in Thailand and Cambodia: An Assessment of the Degree of Underrecognized Disease Burden Based on Reported Cases. Guzman MG, editor. PLoS Negl Trop Dis. Public Library of Science; 2011;5: e996.

doi:10.1371/journal.pntd.0000996

67. Gómez-Dantés H, Willoquet JR. Dengue in the Americas: challenges for prevention and control. Cad Saude Publica. Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz; 2009;25: S19–S31. doi:10.1590/S0102-311X2009001300003

68. Duarte HHP, França EB. Data quality of dengue epidemiological surveillance in Belo Horizonte, Southeastern Brazil. Rev Saude Publica. Faculdade de Saúde Pública da Universidade de São Paulo; 2006;40: 134–142. doi:10.1590/S0034-89102006000100021

69. Wu P-C, Guo H-R, Lung S-C, Lin C-Y, Su H-J. Weather as an effective predictor for occurrence of dengue fever in Taiwan. Acta Trop. Elsevier; 2007;103: 50–57. doi:10.1016/J.ACTATROPICA.2007.05.014

70. Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CAS, Sá Carvalho M, et al. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. Comput Geosci. Pergamon; 2011;37: 371–381. doi:10.1016/J.CAGEO.2010.01.008

71. Johansson MA, Reich NG, Hota A, Brownstein JS, Santillana M, Paulson JA. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico OPEN. Nat Publ Gr. 2016; doi:10.1038/srep33707

72. Reich NG, Lauer SA, Sakrejda K, Iamsirithaworn S, Hinjoy S, Suangtho P, et al.

Challenges in Real-Time Prediction of Infectious Disease: A Case Study of Dengue in Thailand. Scarpino S V., editor. PLoS Negl Trop Dis. Public Library of Science; 2016;10: e0004761. doi:10.1371/journal.pntd.0004761

73. Hii YL, Zhu H, Ng N, Ng LC, Rocklöv J. Forecast of Dengue Incidence Using Temperature and Rainfall. Mutuku F, editor. PLoS Negl Trop Dis. Public Library of Science; 2012;6: e1908. doi:10.1371/journal.pntd.0001908

74. Ramadona AL, Lazuardi L, Hii YL, Holmner Å, Kusnanto H, Rocklöv J. Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data. Moreira LA, editor. PLoS One. Public Library of Science; 2016;11: e0152688. doi:10.1371/journal.pone.0152688

75. Descloux E, Mangeas M, Menkes CE, Lengaigne M, Leroy A, Tehei T, et al. Climate-Based Models for Understanding and Forecasting Dengue Epidemics. Anyamba A, editor. PLoS Negl Trop Dis. Public Library of Science; 2012;6: e1470. doi:10.1371/journal.pntd.0001470

76. Lauer SA, Sakrejda K, Ray EL, Keegan LT, Bi Q, Suangtho P, et al. Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010-2014. Proc Natl Acad Sci U S A. National Academy of Sciences; 2018;115: E2175–E2182. doi:10.1073/pnas.1714457115

77. Juffrie M F DA. EarlyWarning System (EWS) for Dengue in Indonesia and Thailand. J Med Sci (Berkala ilmu Kedokteran). 2009;41. Available: https://jurnal.ugm.ac.id/bik/article/view/2963

78. Bowman LR, Tejeda GS, Coelho GE, Sulaiman LH, Gill BS, McCall PJ, et al. Alarm Variables for Dengue Outbreaks: A Multi-Centre Study in Asia and Latin America. Hsieh Y-H, editor. PLoS One. Public Library of Science; 2016;11: e0157971. doi:10.1371/journal.pone.0157971

79. Centers for Disease Control and Prevention (CDC). Dengue Forecasting Project [Internet]. [cited 18 Feb 2019]. Available: https://predict.cdc.gov/post/5a4fcc3e2c1b1669c22aa261

80. Yamana TK, Kandula S, Shaman J. Superensemble forecasts of dengue outbreaks. J R Soc Interface. The Royal Society; 2016;13: 20160410. doi:10.1098/rsif.2016.0410

81. Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG. Infectious disease prediction with kernel conditional density estimation. Stat Med. John Wiley & Sons, Ltd; 2017;36: 4908–4929. doi:10.1002/sim.7488

82. Johnson LR, Gramacy RB, Cohen J, Mordecai E, Murdock C, Rohr J, et al. Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: a dengue case study. 2017; Available: http://arxiv.org/abs/1702.00261

83. Althouse BM, Ng YY, Cummings DAT. Prediction of Dengue Incidence Using Search Query Surveillance. Crockett RJK, editor. PLoS Negl Trop Dis. Public Library of Science; 2011;5: e1258. doi:10.1371/journal.pntd.0001258

84. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends. Remais J V., editor. PLoS

Negl Trop Dis. Public Library of Science; 2014;8: e2713. doi:10.1371/journal.pntd.0002713

85. Madoff LC, Fisman DN, Kass-Hout T. A New Approach to Monitoring Dengue Activity. Aksoy S, editor. PLoS Negl Trop Dis. Public Library of Science; 2011;5: e1215. doi:10.1371/journal.pntd.0001215

86. Yang S, Kou SC, Lu F, Brownstein JS, Brooke N, Santillana M. Advances in using Internet searches to track dengue. Salathé M, editor. PLOS Comput Biol. Public Library of Science; 2017;13: e1005607. doi:10.1371/journal.pcbi.1005607

87. Anggraeni W, Aristiani L. Using Google Trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia. 2016 International Conference on Information & Communication Technology and Systems (ICTS). IEEE; 2016. pp. 114–118. doi:10.1109/ICTS.2016.7910283

88. Strauss RA, Castro JS, Reintjes R, Torres JR. Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case. Int J Med Inform. Elsevier; 2017;104: 26–30. doi:10.1016/J.IJMEDINF.2017.05.003

89. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. doi:10.1073/pnas.1515373112

90. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using Internet Searches for Influenza Surveillance. Clin Infect Dis. Narnia; 2008;47: 1443–1448. doi:10.1086/593098

91. Valdivia A, López-Alcalde J, Vicente M, Pichiule M, Ruiz M, Ordobas M. Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks – results for 2009-10. Eurosurveillance. European Centre for Disease Prevention and Control; 2010;15: 19621. doi:10.2807/ese.15.29.19621-en

92. Carneiro HA, Mylonakis E. Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. Clin Infect Dis. Narnia; 2009;49: 1557–1564. doi:10.1086/630200

93. Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SC. Using electronic health records and Internet search information for accurate influenza forecasting. doi:10.1186/s12879-017-2424-7

94. Xu Q, Gel YR, Ramirez Ramirez LL, Nezafati K, Zhang Q, Tsui K-L. Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. Cowling BJ, editor. PLoS One. Public Library of Science; 2017;12: e0176690. doi:10.1371/journal.pone.0176690

95. Sharpe JD, Hopkins RS, Cook RL, Striley CW. Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis. JMIR public Heal Surveill. JMIR Public Health and Surveillance; 2016;2: e161. doi:10.2196/publichealth.5901

96. Bardak B, Tan M. Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data. 2015 IEEE 15th International

Conference on Bioinformatics and Bioengineering (BIBE). IEEE; 2015. pp. 1–6. doi:10.1109/BIBE.2015.7367640

97. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. Cowling BJ, editor. PLoS One. Public Library of Science; 2011;6: e23610. doi:10.1371/journal.pone.0023610

98. Lampos V, Miller AC, Crossan S, Stefansen C. Advances in nowcasting influenza-like illness rates using search query logs. Sci Rep. Nature Publishing Group; 2015;5: 12760. doi:10.1038/srep12760

99. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza Forecasting with Google Flu Trends. Viboud C, editor. PLoS One. Public Library of Science; 2013;8: e56176. doi:10.1371/journal.pone.0056176

100. Preis T, Moat HS. Adaptive nowcasting of influenza outbreaks using Google searches. R Soc Open Sci. The Royal Society Publishing; 2014;1: 140095–140095. doi:10.1098/rsos.140095

101. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. Ferguson N, editor. PLoS Comput Biol. Public Library of Science; 2013;9: e1003256. doi:10.1371/journal.pcbi.1003256

102. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS.

Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Comput Biol. 2015;11: 1–15. doi:10.1371/journal.pcbi.1004513

103. Choi H, Varian H. Predicting the Present with Google Trends. Econ Rec. John Wiley & Sons, Ltd (10.1111); 2012;88: 2–9. doi:10.1111/j.1475-4932.2012.00809.x

104. Cho S, Sohn CH, Jo MW, Shin S-Y, Lee JH, Ryoo SM, et al. Correlation between National Influenza Surveillance Data and Google Trends in South Korea. Viboud C, editor. PLoS One. Public Library of Science; 2013;8: e81422. doi:10.1371/journal.pone.0081422

105. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global Disease Monitoring and Forecasting with Wikipedia. Salathé M, editor. PLoS Comput Biol. Public Library of Science; 2014;10: e1003892. doi:10.1371/journal.pcbi.1003892

106. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Med. Public Library of Science; 2008;5: e151. doi:10.1371/journal.pmed.0050151

107. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. J Am Med Informatics Assoc. Narnia; 2008;15: 150–157.

doi:10.1197/jamia.M2544

108. Brownstein JS, Freifeld CC. HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. Euro Surveill. 2007;12: E071129.5. Available: http://www.ncbi.nlm.nih.gov/pubmed/18053570

109. Lyon A, Nunn M, Grossel G, Burgman M. Comparison of Web-Based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap. Transbound Emerg Dis. John Wiley & Sons, Ltd (10.1111); 2012;59: 223–232. doi:10.1111/j.1865-1682.2011.01258.x

110. Liu K, Srinivasan R, Meyers LA. Early Detection of Influenza outbreaks in the United States. 2019; Available: http://arxiv.org/abs/1903.01048

111. Lerman PM. Fitting Segmented Regression Models by Grid Search. Appl Stat. John Wiley & Sons, Ltd (10.1111); 1980;29: 77. doi:10.2307/2346413

112. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. J Mach Learn Res. 2012;13: 281–305. Available: http://www.jmlr.org/papers/v13/bergstra12a.html

113. Bengio Y. Gradient-Based Optimization of Hyperparameters. Neural Comput. MIT Press 238 Main St., Suite 500, Cambridge, MA 02142-1046 USA journals-info@mit.edu ; 2000;12: 1889–1900. doi:10.1162/089976600300015187

114. Simon D. Evolutionary optimization algorithms [Internet]. Available: https://books.google.com/books?hl=en&lr=&id=gwUwIEPqk30C&oi=fnd&pg=PP

1&dq=Evolutionary+optimization&ots=GLo5ApMeh9&sig=t6M1J-AUtsgKECyTfiCe8ltCb5s#v=onepage&q=Evolutionary optimization&f=false

115. Fogel DB. An introduction to simulated evolutionary optimization. IEEE Trans neural networks. 1994;5: 3–14. Available: http://l.academicdirect.org/Horticulture/GAs/Refs/_other_Fogel/Fogel_1994_Evolution.pdf

116. Lizotte DJ. Practical Bayesian optimization [Internet]. Library and Archives Canada. 2008. Available: https://dl.acm.org/citation.cfm?id=1626686

117. Bergstra, James S., Rémi B, Yoshua Bengio BK. Algorithms for hyper-parameter optimization. Advances in neural information processing systems. 2011. pp. 2546–2554. Available: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimizat

118. Lowry CA, Woodall WH, Champ CW, Rigdon SE. A Multivariate Exponentially Weighted Moving Average Control Chart A Multivariate Exponentially Weighted Moving Average Control Chart. Source: Technometrics. 1992;3434: 46–5346. Available: http://www.jstor.org/stable/1269551

119. H. Hotelling. Multivariable Quality Control—Illustrated by the Air Testing of Sample Bombsight. Techniques of Statistical Analysis. New York: McGraw Hill; 1947. pp. 110–122.

120. Hyperopt. Distributed Asynchronous Hyper-parameter Optimization [Internet]. [cited 28 Oct 2018]. Available: http://hyperopt.github.io/hyperopt/

121. Secretaría de Salud. Histórico Boletín Epidemiológico [Internet]. [cited 20 Sep 2018]. Available: https://www.gob.mx/salud/acciones-y-programas/historico-boletin-epidemiologico

122. National Oceanic and Atmospheric Administration (NOAA) [Internet]. Available: https://www.noaa.gov/

123. Brady OJ, Johansson MA, Guerra CA, Bhatt S, Golding N, Pigott DM, et al. Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. Parasit Vectors. BioMed Central; 2013;6: 351. doi:10.1186/1756-3305-6-351

124. Juliano SA, O'Meara GF, Morrill JR, Cutwa MM. Desiccation and thermal tolerance of eggs and the coexistence of competing mosquitoes. Oecologia. Springer Berlin Heidelberg; 2002;130: 458–469. doi:10.1007/s004420100811

125. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 2013;496: 504–507. doi:10.1038/nature12060

126. Scott TW, Morrison AC, Lorenz LH, Clark GG, Strickman D, Kittayapong P, et al. Longitudinal Studies of *Aedes aegypti* (Diptera: Culicidae) in Thailand and Puerto Rico: Population Dynamics. J Med Entomol. Narnia; 2000;37: 77–88. doi:10.1603/0022-2585-37.1.77

127. Burke DS, Scott RM, Johnson DE, Nisalak A. A Prospective Study of Dengue Infections in Bangkok. Am J Trop Med Hyg. The American Society of Tropical

Medicine and Hygiene; 1988;38: 172–180. doi:10.4269/ajtmh.1988.38.172

128. Promprou S, Jaroensutasinee M, Jaroensutasinee K. Climatic Factors Affecting Dengue Haemorrhagic Fever Incidence in Southern Thailand [Internet]. Dengue Bulletin. 2005. Available: https://apps.who.int/iris/bitstream/handle/10665/164135/dbv29p41.pdf

129. Thu, Hlaing Myat, Khin Mar Aye ST. The effect of temperature and humidity on dengue virus propagation in Aedes aegypti mosquitos. Southeast Asian J Trop Med Public Heal. 1998;29: 280–284. Available: http://www.thaiscience.info/journals/Article/TMPH/10726187.pdf

130. Global Historical Climatology Network (GHCN) [Internet]. [cited 28 May 2019]. Available: https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-ghcn

131. Integrated Surface Database (ISD) [Internet]. [cited 28 May 2019]. Available: https://www.ncdc.noaa.gov/isd

132. National Centers for Environmental Prediction [Internet]. [cited 28 May 2019]. Available: https://www.ncep.noaa.gov/

133. Lowe R, Stewart-Ibarra AM, Petrova D, García-Díez M, Borbor-Cordova MJ, Mejía R, et al. Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador. Lancet Planet Heal. Elsevier; 2017;1: e142–e151. doi:10.1016/S2542-5196(17)30064-5

134. Lowe R, Gasparrini A, Van Meerbeeck CJ, Lippi CA, Mahon R, Trotman AR, et al. Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. Thomson M, editor. PLOS Med. Public Library of Science; 2018;15: e1002613. doi:10.1371/journal.pmed.1002613

135. Cazelles B, Chavez M, McMichael AJ, Hales S. Nonstationary Influence of El Niño on the Synchronous Dengue Epidemics in Thailand. Pascual M, editor. PLoS Med. Public Library of Science; 2005;2: e106. doi:10.1371/journal.pmed.0020106

136. Fuller DO, Troyo A, Beier JC. El Niño Southern Oscillation and vegetation dynamics as predictors of dengue fever cases in Costa Rica. Environ Res Lett. IOP Publishing; 2009;4: 014011. doi:10.1088/1748-9326/4/1/014011

137. Earnest A, Tan SB, Wilder-Smith A. Meteorological factors and El Niño Southern Oscillation are independently associated with dengue infections. Epidemiol Infect. Cambridge University Press; 2012;140: 1244–1251. doi:10.1017/S095026881100183X

138. Hu W, Clements A, Williams G, Tong S. Dengue fever and El Nino/Southern Oscillation in Queensland, Australia: a time series predictive model. Occup Environ Med. BMJ Publishing Group Ltd; 2010;67: 307–11. doi:10.1136/oem.2008.044966

139. Gagnon A, Bush A, Smoyer-Tomic K. Dengue epidemics and the El Niño Southern Oscillation. Clim Res. 2001;19: 35–43. doi:10.3354/cr019035

140. Tipayamongkholgul M, Fang C-T, Klinchan S, Liu C-M, King C-C. Effects of the

El Niño-Southern Oscillation on dengue epidemics in Thailand, 1996-2005. BMC Public Health. BioMed Central; 2009;9: 422. doi:10.1186/1471-2458-9-422

141. Sehgal R. Dengue fever and El Niño. Lancet (London, England). Elsevier; 1997;349: 729. doi:10.1016/S0140-6736(05)60169-9

142. Google Trends [Internet]. [cited 28 May 2019]. Available: https://trends.google.com/trends/?geo=US

143. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science (80-). American Association for the Advancement of Science; 2014;343: 1203–1205.

144. Butler D. When Google got flu wrong. Nature. 2013;494: 155–156. doi:10.1038/494155a

145. Yeh S-W, Kug J-S, Dewitte B, Kwon M-H, Kirtman BP, Jin F-F. El Niño in a changing climate. Nature. Nature Publishing Group; 2009;461: 511–514. doi:10.1038/nature08316

146. Johansson MA, Dominici F, Glass GE. Local and Global Effects of Climate on Dengue Transmission in Puerto Rico. Massad E, editor. PLoS Negl Trop Dis. Public Library of Science; 2009;3: e382. doi:10.1371/journal.pntd.0000382

147. Wikipedia. Iquitos [Internet]. [cited 28 May 2019]. Available: https://en.wikipedia.org/wiki/Iquitos

148. Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG, Evan Lowell Ray C.

Infectious disease prediction with kernel conditional density estimation. 2017; doi:10.1002/sim.7488

149. Mooney, CZ., Robert D. Duval RD. Bootstrapping: A Nonparametric Approach to Statistical Inference [Internet]. Sage; 1993. Available: https://books.google.com/books?hl=en&lr=&id=ZxaRC4I2z6sC&oi=fnd&pg=PP6&dq=bootstrapping&ots=oQl8FscZuI&sig=vLpQ2sDwHqb_TyEeFEVpnK3iHKc#v=onepage&q=bootstrapping&f=false

150. Lu FS, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, et al. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. JMIR public Heal Surveill. JMIR Public Health and Surveillance; 2018;4: e4. doi:10.2196/publichealth.8950

151. Xu H-Y, Fu X, Lee LKH, Ma S, Goh KT, Wong J, et al. Statistical Modeling Reveals the Effect of Absolute Humidity on Dengue in Singapore. Barrera R, editor. PLoS Negl Trop Dis. Public Library of Science; 2014;8: e2805. doi:10.1371/journal.pntd.0002805

152. World Health Organization. Ebola situation reports (17 February 2016) [Internet]. [cited 4 Jun 2019]. Available: https://apps.who.int/iris/bitstream/handle/10665/204418/ebolasitrep_17Feb2016_eng.pdf?sequence=1

153. Chretien J-P, Riley S, George DB. Mathematical modeling of the West Africa Ebola epidemic. Elife. 2015;4: 1689–1699. doi:10.7554/eLife.09186

154. Meltzer MI, Atkins CY, Santibanez S, Knust B, Petersen BW, Ervin ED, et al. Estimating the future number of cases in the Ebola epidemic--Liberia and Sierra Leone, 2014-2015. MMWR Suppl. 2014;63: 1–14. Available: http://www.ncbi.nlm.nih.gov/pubmed/25254986

155. Gerardo C, Cécile V, James M. Hyman LS. The Western Africa Ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. arxiv. 2014;

156. Chowell G, Viboud C, Simonsen L, Moghadas S. Characterizing the reproduction number of epidemics with early sub-exponential growth dynamics. 2016.

157. Merler S, Ajelli M, Fumanelli L, Gomes MFC, Piontti AP, Rossi L, et al. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. Lancet Infect Dis. Elsevier; 2015;15: 204–211. doi:10.1016/S1473-3099(14)71074-6

158. Scarpino S V., Iamarino A, Wells C, Yamin D, Ndeffo-Mbah M, Wenzel NS, et al. Epidemiological and Viral Genomic Sequence Analysis of the 2014 Ebola Outbreak Reveals Clustered Transmission. Clin Infect Dis. 2014; 1–4. doi:10.1093/cid/ciu1131

159. Matanock A, Arwady MA, Ayscue P, Forrester JD, Gaddis B, Hunter JC, et al. Ebola virus disease cases among health care workers not working in Ebola treatment units--Liberia, June-August, 2014. MMWR Morb Mortal Wkly Rep.

Centers for Disease Control and Prevention; 2014;63: 1077–81. Available: http://www.ncbi.nlm.nih.gov/pubmed/25412067

160. Nyenswah T, Massaquoi M, Gbanya MZ, Fallah M, Amegashie F, Kenta A, et al. Initiation of a ring approach to infection prevention and control at non-Ebola health care facilities - Liberia, January-February 2015. MMWR Morb Mortal Wkly Rep. Centers for Disease Control and Prevention; 2015;64: 505–8. Available: http://www.ncbi.nlm.nih.gov/pubmed/25974636

161. Meyers LA, Newman MEJ, Pourbohloul B. Predicting epidemics on directed contact networks. J Theor Biol. 2006;240: 400–18. doi:10.1016/j.jtbi.2005.10.004

162. Casillas AM, Nyamathi AM, Sosa A, Wilder CL, Sands H. A Current Review of Ebola Virus: Pathogenesis, Clinical Presentation, and Diagnostic Assessment. Biol Res Nurs. SAGE Publications; 2003;4: 268–275. doi:10.1177/1099800403252603

163. Meyers LA, Pourbohloul B, Newman MEJ, Skowronski DM, Brunham RC. Network theory and SARS: predicting outbreak diversity. J Theor Biol. 2005;232: 71–81. doi:10.1016/j.jtbi.2004.07.026

164. Bansal S, Grenfell BT, Meyers L a. When individual behaviour matters: homogeneous and network models in epidemiology. J R Soc Interface. 2007;4: 879–891. doi:10.1098/rsif.2007.1100

165. Volz E, Meyers LA. Susceptible-infected-recovered epidemics in dynamic contact networks. Proc Biol Sci. 2007;274: 2925–33. doi:10.1098/rspb.2007.1159

166. Herrera JL, Srinivasan R, Brownstein JS, Galvani AP, Meyers LA. Disease Surveillance on Complex Social Networks. Salathé M, editor. PLOS Comput Biol. Public Library of Science; 2016;12: e1004928. doi:10.1371/journal.pcbi.1004928

167. Volz EM, Miller JC, Galvani A, Ancel Meyers L. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. PLoS Comput Biol. 2011;7: e1002042. doi:10.1371/journal.pcbi.1002042

168. Miller JC, Slim AC, Volz EM. Edge-based compartmental modelling for infectious disease spread. 2011;

169. Miller JC, Volz EM. Edge-Based Compartmental Modeling for Infectious Disease Spread Part III: Disease and Population Structure. 2011; 1–17. Available: http://arxiv.org/abs/1106.6344

170. Miller JC. Epidemics on networks with large initial conditions or changing structure. PLoS One. 2014;9: 1–9. doi:10.1371/journal.pone.0101421

171. WHO Ebola Response Team. Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections. Engl New J Med. 2014; 1–15. doi:10.1056/NEJMoa1411100

172. The SciPy community. SciPy [Internet]. [cited 4 Jun 2019]. Available: https://docs.scipy.org/doc/scipy-1.2.1/reference/

173. World Health Organization. Global Health Workforce Statistics [Internet]. [cited 4 Jun 2019]. Available: http://apps.who.int/gho/data/node.main.HWFGRP?lang=en

174. Newville M, Stensitzki T, et al. Non-Linear Least-Squares Minimization and Curve-Fitting for Python [Internet]. 2018. Available: http://cars9.uchicago.edu/software/python/lmfit/lmfit.pdf

175. Karrer B, Newman MEJ. Random graphs containing arbitrary distributions of subgraphs. Phys Rev E - Stat Nonlinear, Soft Matter Phys. 2010;82: 1–12. doi:10.1103/PhysRevE.82.066118

Vita

Kai Liu was born and grew up in Jiaozhou, China, where he attended primary school, middle school and high school. After graduating from Jiaozhou Experimental High School in 2007, he attended Huazhong Agricultural University, and received a Bachelor of Science in Biotechnology in 2011 and a Master of Science in Microbiology in 2013. From there, he decided to come to Cellular and Molecular Biology program at the University of Texas at Austin, and joined Meyers Lab for his doctoral study in 2014 after three lab rotations.

Permanent email: liukaifun@gmail.com

This dissertation was typed by the author.