Journal of Biomedical Informatics 73 (2017) 171–181

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A Bayesian system to detect and characterize overlapping outbreaks

John M. Aronis [a,*], Nicholas E. Millett [a], Michael M. Wagner [a,b], Fuchiang Tsui [a,b], Ye Ye [a,b], Jeffrey P. Ferraro [c,d], Peter J. Haug [c,d], Per H. Gesteland [c,d,e], Gregory F. Cooper [a,b]

[a] Real-time Outbreak and Disease Surveillance Laboratory, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
[b] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA
[c] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
[d] Intermountain Healthcare, Salt Lake City, UT, USA
[e] Department of Pediatrics, University of Utah, Salt Lake City, UT, USA

ARTICLE INFO

Article history:
Received 28 September 2016
Revised 4 July 2017
Accepted 4 August 2017
Available online 7 August 2017

Keywords:
Influenza
Outbreak detection
Outbreak characterization
Bayesian modeling

ABSTRACT

Outbreaks of infectious diseases such as influenza are a significant threat to human health. Because there are different strains of influenza which can cause independent outbreaks, and influenza can affect demographic groups at different rates and times, there is a need to recognize and characterize multiple outbreaks of influenza. This paper describes a Bayesian system that uses data from emergency department patient care reports to create epidemiological models of overlapping outbreaks of influenza. Clinical findings are extracted from patient care reports using natural language processing. These findings are analyzed by a case detection system to create disease likelihoods that are passed to a multiple outbreak detection system. We evaluated the system using real and simulated outbreaks. The results show that this approach can recognize and characterize overlapping outbreaks of influenza. We describe several extensions that appear promising.

1. Introduction

Influenza is a contagious viral disease that spreads around the world in seasonal epidemics. It causes 3–5 million severe illnesses and 250,000–500,000 deaths annually. It also causes considerable health care costs and lost productivity [1]. There is a clear need for accurate, timely predictions of influenza outbreaks for hospitals, public health officials, and business [2].

This paper describes a *Multiple Outbreak Detection System* (MODS) that detects and characterizes overlapping outbreaks of influenza. MODS is part of a framework for disease surveillance developed by our group [3–5]. In this framework, a natural language processing system extracts signs and symptoms from the full text of emergency department (ED) patient care reports. These extracted features are combined with laboratory values and passed to a *Case Detection System* (CDS) that infers a probability distribution over the diseases each patient may have. For each patient, this distribution is expressed as a set of likelihoods of the patient's data. MODS searches a space of outbreak models and scores each model according to the probability of the findings of each patient in the ED given the model. Models are weighted and combined to make predictions. See Fig. 1.

There has been significant prior work in the detection and characterization of outbreaks of influenza [6–10]. The work reported here differs from much of the prior work in three important ways. First, we address the problem of detecting and characterizing multiple, overlapping outbreaks. These are commonly due to separate type A and type B infections, but may also be caused by multiple introductions of the same influenza strain, the spread of an influenza strain in different demographic groups, or different strains of influenza [1,11–14]. Second, we do not rely on simple counts. Instead, for each patient, CDS produces a set of likelihoods of that patient's evidence given the diseases he or she may have. This is more robust than a posterior probability. While the posterior probability of influenza given fever and cough in a particular patient will change if it is summer, the peak of an influenza outbreak, or the days following an anthrax attack [15], the likelihood of fever and cough given influenza is relatively stable. Finally, we explicitly account for *Non-Influenza Influenza-Like Illnesses* (NI-ILI) which are clinically similar to influenza and can also display outbreak activity.

2. Material and methods

2.1. Evidence

As stated above, MODS is part of an end-to-end system that starts with the full text of emergency department patient care

* Corresponding author.
E-mail address: jma18@pitt.edu (J.M. Aronis).

http://dx.doi.org/10.1016/j.jbi.2017.08.003
1532-0464/

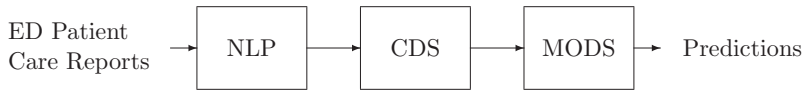

Fig. 1. End-to-end framework for outbreak detection and characterization.

reports. These reports contain a record of the patient assessment, past medical history, and history of present illness. To model outbreaks of influenza, these can be more informative than just chief complaints [16]. In this section, we describe how these reports are processed to support outbreak detection and characterization.

The text of each patient care report is parsed to extract 79 findings (such as *abdominal pain*, *cough*, and *fever*) that were deemed by experts to be relevant to the diagnosis of influenza and influenza-like illnesses. Thus, each report is converted into a set of 79 features. Each feature can take the values *present* (the corresponding finding is explicitly mentioned, *e.g.* "the patient is cyanotic."), *absent* (the finding is denied, *e.g.* "the patient denies chest pain."), or *missing* (the finding is not mentioned in the text). Four pre-coded features, *age*, *respiratory panel ordered*, *laboratory result available*, and *laboratory positive influenza* are then added to the text-based 79 findings. Thus, each report is converted to a set of 83 features. These features are listed in Table 1.

Table 1
Features extracted from reports with NLP plus precoded findings. Precoded findings are in italics. "AC" or "SLC" indicate if finding was used in CDS for Allegheny County or Salt Lake City.

abdominal distress	lab testing ordered (influenza) (AC)
abdominal pain	lab testing ordered (rsv) (SLC)
abdominal stress	lab testing ordered (panel) (SLC)
abdominal tenderness	malaise
abnormal chest radiograph findings	myalgia (AC)
acute onset	nasal flaring
anorexia	nausea
apnea (SLC)	nonproductive cough (AC)
arthralgia	non-specific cough (other cough) (AC, SLC)
barking cough	other abnormal breath sounds
bilateral acute conjunctivitis	other pneumonia (AC)
bronchiolitis (AC)	paroxysmal cough
bronchitis	pharyngitis diagnosis
cervical lymphadenopathy	pharyngitis on exam
chest pain	poor feeding
chest wall retractions	poor antipyretics response
chills	productive cough
conjunctivitis	rales
crackles	reported fever (AC, SLC)
croup	respiratory distress (SLC)
cyanosis	rhonchi
decreased activity	rigor
diarrhea	runny nose
dyspnea	seizure
grunting	sore throat
headache	staccato cough
hemoptysis	streptococcal pharyngitis
highest measured temperature (SLC)	stridor
hoarseness	stuffy nose
hypoxemia (AC, SLC)	tachypnea (SLC)
ill-appearing (AC)	toxic appearance
infiltrate	upper respiratory infection
influenza-like illness (AC)	viral pneumonia
lab order (nasal swab) (AC, SLC)	viral syndrome (AC)
lab positive adenovirus	vomiting
lab positive enterovirus	weakness or fatigue
lab positive hmpv	wheezing
lab positive influenza (AC, SLC)	*age* (AC, SLC)
lab positive parainfluenza	*respiratory panel ordered* (AC, SLC)
lab positive rhinovirus	*laboratory result available* (AC, SLC)
lab positive RSV	*laboratory positive influenza* (AC, SLC)
lab positive strep A	

Two site-specific case detection systems were built, one for Allegheny County and one for Salt Lake City. Each case detection system was automatically constructed from training data using a machine learning algorithm that performs feature selection followed by the construction of a Bayesian network using the selected features. Based on this process, the Bayesian network for Allegheny County contains 17 of the original 83 features, and the network for Salt Lake City contains 15. (The selected features are marked in Table 1 with "AC" or "SLC.") Each case detection system is a Bayesian network that represents the conditional probability distribution of findings and disease state (*flu*, *NI-ILI*, or *other*). This distribution can provide the probability of that patient's findings given their assumed disease state. That is, $P(E(p,d)|flu)$, $P(E(p,d)|NI\text{-}ILI)$, and $P(E(p,d)|other)$ where $E(p,d)$ denotes the findings for patient p on day d. (Note that $E(p,d)$ consists of 17 findings for each patient in Allegheny County and 15 findings for each patient in Salt Lake City due to feature selection.) These likelihoods—$P(E(p,d)|flu)$, $P(E(p,d)|NI\text{-}ILI)$, and $P(E(p,d)|other)$ for each patient—are passed to MODS.

Note that MODS does not classify patients as *flu*, *NI-ILI*, or *other*, and does not use simple evidence counts. Rather, it derives likelihoods that represent the extend to which a patient's findings are more or less likely given *flu*, *NI-ILI*, or *other*. This is in keeping with our Bayesian approach which seeks to build a model that best explains the data.

To summarize, we start with the full text of ED patient care reports. We use NLP to extract 79 features from each report, add four additional pre-coded features (including the result of a laboratory test), to represent each patient as a set of 83 features. We then use a case detection system to produce three likelihoods ($P(E(p,d)|flu), P(E(p,d)|NI\text{-}ILI)$, and $P(E(p,d)|other)$) for each patient. Thus, the dataset given to MODS after preprocessing consists of three likelihoods for each patient. Additional details are provided in Appendix A.

Although MODS does not categorically classify patients, we can count ILI (*influenza-like illness*) patients by selecting cases with findings that include *fever and (cough or sore throat)*. Although MODS does not use this information, it can provide a convenient way to visualize the data. Fig. 2 shows the number of emergency department ILI cases counted in this way for Allegheny County during the 2009–2010 influenza season, and Fig. 3 shows the number for Sale Lake City during the 2010–2011 influenza season. We include this plot on graphs in this paper.

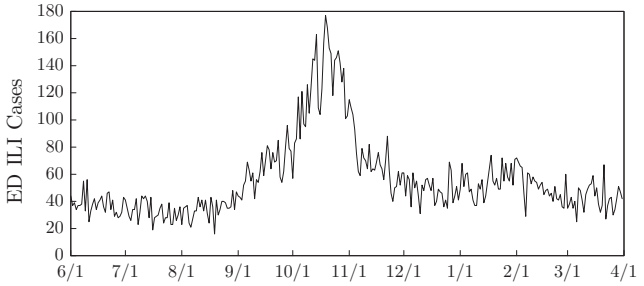

Fig. 2. ED ILI cases for Allegheny County 2009–2010 influenza season.

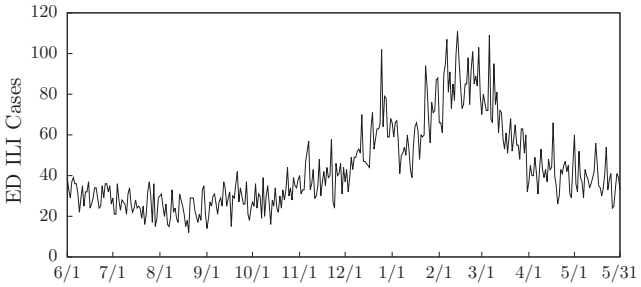

Fig. 3. ED ILI cases for Salt Lake City 2010–2011 influenza season.

2.2. Modeling outbreaks

An outbreak of influenza is a chain of infections that starts when a few individuals infect the general population, the infection spreads evenly and steadily through the population until the number of infected individuals reaches a peak, then decreases and finally trails off. More generally, an outbreak may be caused by different strains of influenza that are introduced into different demographic groups at different times. That is, multiple chains of infection, each with its own start time.

2.2.1. SEIR models

A SEIR model [17] divides the population into four compartments: *Susceptible* individuals who have no immunity to the disease, *Exposed and Infected* individuals who are not yet contagious, *Infectious* individuals who have the disease and can transmit it, and *Recovered* individuals who are immune. When an outbreak begins, every individual is assigned to one compartment and move from compartment to compartment according to difference equations.

A SEIR model is defined by seven parameters. R_0 (the *Basic Reproduction Number*) is the average number of people an infectious person would infect if they were introduced into a population of totally susceptible individuals, *Latent Period* is the average time from initial contact with an infected person to when a person becomes infectious, and *Infectious Period* is the average time an individual is infectious and can transmit the disease. Also, *Initial Susceptible*, *Initial Exposed*, *Initial Infectious*, and *Initial Recovered* are the initial numbers of people assigned to the corresponding compartments on the first day of the outbreak when the population is initially infected.

Given the parameters of a SEIR model, we can compute the number of people in the *Susceptible*, *Exposed and Infected*, *Infectious*, and *Recovered* compartments each day using a set of difference equations [17].

2.2.2. Modeling multiple outbreaks of influenza

MODS includes three kinds of influenza models corresponding to no outbreak, one chain of infections, and two chains of infections:

- A *Zero Outbreak Model* is defined by b, a baseline level of infectious influenza cases in the general population. We denote the zero outbreak model with $\langle b \rangle$.
- A *Single Outbreak Model* is defined by a baseline level of infectious influenza cases in the general population, a start day, and a SEIR model. Given a baseline level of infectious influenza cases b, a start day s, and a vector of SEIR parameters $\overline{\lambda}$, we denote the single outbreak model with $\langle b, s, \overline{\lambda} \rangle$. Note that $\langle b, s, \overline{\lambda} \rangle$ has a total of nine parameters: a baseline level of influenza, the start day, and the vector $\overline{\lambda}$ of seven SEIR parameters.
- A *Double Outbreak Model* is defined by a baseline level of infectious influenza cases in the general population, a start day and SEIR parameters for the first outbreak, and a start day and SEIR parameters for the second outbreak. Given a baseline level of infectious influenza cases b, a start day s and a vector of SEIR parameters $\overline{\lambda}$, and a second start day t and vector of SEIR parameters $\overline{\gamma}$, we denote the double outbreak model with $\langle b, s, \overline{\lambda}, t, \overline{\gamma} \rangle$. Note that $\langle b, s, \overline{\lambda}, t, \overline{\gamma} \rangle$ has a total of seventeen parameters: the baseline level of infectious influenza cases, two start days, and two vectors of seven SEIR parameters each.

A single outbreak model $\langle b, s, \overline{\lambda} \rangle$ is intended to model the introduction on day s of an infection that spreads according to the SEIR model defined by the parameters $\overline{\lambda}$. Likewise, a double outbreak model $\langle b, s, \overline{\lambda}, t, \overline{\gamma} \rangle$ is intended to model the introduction on day s of an infection that spreads according to the SEIR model defined by the parameters $\overline{\lambda}$, and a second infection introduced on day t that spreads according to the SEIR model defined by the parameters $\overline{\gamma}$. We often denote an influenza model by $\mathcal{I}$ rather than write out $\langle b \rangle$, $\langle b, s, \overline{\lambda} \rangle$, or $\langle b, s, \overline{\lambda}, t, \overline{\gamma} \rangle$.

If $\mathcal{I}$ is a zero, single, or double outbreak model of influenza, it determines $\mathcal{I}(d)$, the number of infectious influenza cases in the general population on day d, in the following ways:

- If $\mathcal{I}$ is a zero outbreak model $\langle b \rangle$, then $\mathcal{I}(d) = b$ is constant.
- If $\mathcal{I}$ is a single outbreak model $\langle b, s, \overline{\lambda} \rangle$, then $\mathcal{I}(d)$ is b plus the number of infectious influenza cases in the general population on day d predicted by the SEIR model $\overline{\lambda}$ starting on the start day s.
- If $\mathcal{I}$ is a double outbreak model $\langle b, s, \overline{\lambda}, t, \overline{\gamma} \rangle$, then $\mathcal{I}(d)$ is b, plus the number of infectious influenza cases in the general population on day d predicted by the SEIR model $\overline{\lambda}$ starting on its start day s, plus the number of infectious cases in the general population on day d predicted by the SEIR model $\overline{\gamma}$ starting on its start day t.

That is, we compute $\mathcal{I}(d)$ by adding b, plus the number of infectious influenza cases predicted by the first SEIR model (if there is one), plus the number of infectious influenza cases predicted by the second SEIR model (if there is one). The number of infectious influenza cases predicted by each SEIR model is the number of cases in its *Infectious* compartment, which can be computed with a set of difference equations, as mentioned above. Note that we make the simplifying assumption that there is a constant, baseline level of infectious influenza cases (b) in the general population.

2.2.3. Influenza in the general population and the emergency department

MODS searches a space of outbreak models. For each model $\mathcal{I}$ considered, $\mathcal{I}(d)$ is the number of infectious influenza cases in the general population on day d predicted by the model. However, only a fraction of these cases go to an ED. We rely on the scaling factor θ which is the probability that someone who is infectious with influenza will actually go to a emergency department (ED) on any given day. In general, we do not know θ exactly, so we integrate over a range of values. However, we do assume it is constant for the duration of an outbreak.

2.2.4. Accounting for non-influenza influenza-like illness

Clinically, it is difficult or impossible to distinguish influenza from a variety of *Influenza-Like Illnesses* (ILI). Most kinds of ILI are viral—such as respiratory syncytial virus (RSV) and parainfluenza—but some other diseases and toxic exposures also cause influenza-like symptoms. In most years, influenza is present in

only a minority of ILI cases, but may account for 50–60% of ILI cases at the peak of an epidemic [18]. Because some kinds of ILI are contagious and exhibit outbreak activity [17], it is important to distinguish them from true influenza in order to accurately predict and characterize outbreaks of influenza. We rely on the use of laboratory tests, such as the polymerase chain reaction (PCR) test, to distinguish influenza from NI-ILI. Because laboratory tests are ordered for relatively few patients, we have developed a technical approach to calculate $\mathcal{N}(d)$, the expected number of NI-ILI cases in the ED on each day d. The details of our method are described in Appendix B.

2.2.5. Evaluating models

Let E be all of the evidence available to the system, $E(d)$ the evidence from day d, $E(1:d)$ the evidence from days 1 through d, and $E(p,d)$ the evidence from patient p on day d. CDS represents the joint probability distribution of findings and diseases, so we can find $P(E(p,d)|flu), P(E(p,d)|NI\text{-}ILI)$, $P(E(p,d)|other)$, the probability of a patient's findings given influenza, NI-ILI, or some other diagnosis (such as appendicitis and trauma).

We evaluate a model $\mathcal{I}$ by how well it explains the data—that is, the probability of the data given $\mathcal{I}$. Since $\mathcal{I}$ predicts the number of influenza cases in the population—but the data is drawn from hospitals—θ (the fraction of people infectious with influenza who will go to an ED) plays a role. We do not know θ precisely, but our belief in its value is uniform over an interval $[\theta_l, \theta_u]$. Thus, given data $E(1:d)$, we evaluate a model $\mathcal{I}$ with:

$$P(E(1:d)|\mathcal{I}) = \frac{1}{(\theta_u - \theta_l)} \int_{\theta_l}^{\theta_u} P(E(1:d)|\mathcal{I}, \theta)\, d\theta \tag{1}$$

We also need to account for the effect of NI-ILI on the data. Appendix B shows how to construct a function $\mathcal{N}(d)$ that estimates the number of NI-ILI patients in the ED each day d and shows:

$$P(E(1:d)|\mathcal{I}, \theta) = \prod_{d'=1}^{d} P(E(d')|\mathcal{I}(d'), \mathcal{N}(d'), \theta) \tag{2}$$

$\mathcal{I}(d')$ is the number of cases of infectious influenza in the general population predicted by the model $\mathcal{I}$ on day d'. Section 2.2.2 describes how $\mathcal{I}(d')$ is computed. $\mathcal{N}(d')$ is the number of NI-ILI patients in the ED given by the model of NI-ILI. If we assume that given $\mathcal{I}, \mathcal{N}$, and θ the probability of each patient's evidence is independent of the other patients, we have:

$$P(E(d)|\mathcal{I}(d), \mathcal{N}(d), \theta) = \prod_{p \in pts(d)} P(E(p,d)|\mathcal{I}(d), \mathcal{N}(d), \theta) \tag{3}$$

Putting all of this together gives:

$$P(E(1:d)|\mathcal{I}) = \frac{1}{(\theta_u - \theta_l)} \int_{\theta_l}^{\theta_u} \prod_{d'=1}^{d} \prod_{p \in pts(d')} P(E(p,d')|\mathcal{I}(d'), \mathcal{N}(d'), \theta)\, d\theta \tag{4}$$

We estimate the probability that an arbitrary ED patient has influenza as $\theta \cdot \mathcal{I}(d)/|pts(d)|$ and the probability he has NI-ILI as $\mathcal{N}(d)/|pts(d)|$ (where $pts(d)$ is the number of ED patients on day d), so:

$$\begin{aligned} P(E(p,d)|\mathcal{I}(d), \mathcal{N}(d), \theta) = {} & P(E(p,d)|flu)(\theta \cdot \mathcal{I}(d)/|pts(d)|) \\ & + P(E(p,d)|NI\text{-}ILI)(\mathcal{N}(d)/|pts(d)|) \\ & + P(E(p,d)|other)((|pts(d)| - (\theta \cdot \mathcal{I}(d) \\ & + \mathcal{N}(d)))/|pts(d)|) \end{aligned} \tag{5}$$

Although Eq. (4) is a closed-form equation involving only the variable θ, we have no practical way to integrate it. However, given values for $\mathcal{I}(d)$, $\mathcal{N}(d)$ and θ, we can evaluate the expression inside the integral, so we approximate the integral with:

$$P(E(1:d)|\mathcal{I}) \approx \frac{1}{(\theta_u - \theta_l)} \frac{1}{n} \sum_{i=1}^{n} \prod_{d'=1}^{d} \prod_{p \in pts(d')} P(E(p,d')|\mathcal{I}(d'), \mathcal{N}(d'), \theta_i) \tag{6}$$

where $\theta_1, \ldots, \theta_n$ are randomly and uniformly drawn from $[\theta_l, \theta_u]$.

2.3. Detection and characterization

2.3.1. Detecting outbreaks

Let $\mathbf{Z}$, $\mathbf{S}$, and $\mathbf{D}$ be the propositions that zero, one (single), or two (double) outbreaks of influenza will occur during the year (365 days starting from day 1). Let $\mathbf{Z}_d$, $\mathbf{S}_d$, and $\mathbf{D}_d$ be the propositions that zero, one, or two outbreaks of influenza start on or before day d. We want to find $P(\mathbf{Z}_d|E(1:d)), P(\mathbf{S}_d|E(1:d))$, and $P(\mathbf{D}_d|E(1:d))$. That is, the probability that zero, one, or two influenza outbreaks have already begun given the evidence so far. Using Bayes Rule we obtain:

$$P(\mathbf{Z}_d|E(1:d)) = P(E(1:d)|\mathbf{Z}_d)P(\mathbf{Z}_d)/P(E(1:d)) \tag{7}$$

$$P(\mathbf{S}_d|E(1:d)) = P(E(1:d)|\mathbf{S}_d)P(\mathbf{S}_d)/P(E(1:d)) \tag{8}$$

$$P(\mathbf{D}_d|E(1:d)) = P(E(1:d)|\mathbf{D}_d)P(\mathbf{D}_d)/P(E(1:d)) \tag{9}$$

where $P(E(1:d))$ is the standard normalizing term.

First, we will derive $P(\mathbf{Z}_d), P(\mathbf{S}_d)$, and $P(\mathbf{D}_d)$. $\mathbf{Z}_d$ claims that no outbreak will start on or before day d, but zero, one, or two may start later. If $f(s)$ is the probability that an outbreak will start on day s, then:

$$P(\mathbf{Z}_d) = P(\mathbf{Z}_d|\mathbf{Z})P(\mathbf{Z}) + P(\mathbf{Z}_d|\mathbf{S})P(\mathbf{S}) + P(\mathbf{Z}_d|\mathbf{D})P(\mathbf{D}) \tag{10}$$

$$= P(\mathbf{Z}) + \left(\sum_{s=(D+1)}^{365} f(s)\right) P(\mathbf{S}) + \left(\sum_{s=(D+1)}^{365} f(s)\right)^2 P(\mathbf{D}) \tag{11}$$

The first term expresses the fact that the probability that no outbreaks start on or before day d given no outbreaks start at all is one. The second term expresses the fact that the probability that no outbreaks start on or before day d given one outbreak will occur during the year is the probability that outbreak begins after day d. The third term expresses the fact that the probability that no outbreaks start on or before day d given two outbreaks will occur during the year is the probability that both outbreaks begin after day d.

$\mathbf{D}_d$ claims that two outbreaks of influenza start on or before day d. This logically rules out $\mathbf{Z}$ and $\mathbf{S}$. Since there are exactly two outbreaks and both begin on or before day d we have:

$$P(\mathbf{D}_d) = P(\mathbf{D}_d|\mathbf{Z})P(\mathbf{Z}) + P(\mathbf{D}_d|\mathbf{S})P(\mathbf{S}) + P(\mathbf{D}_d|\mathbf{D})P(\mathbf{D}) \tag{12}$$

$$= \left(\sum_{s=1}^{d} f(s)\right)^2 P(\mathbf{D}) \tag{13}$$

Finally, by simple probability theory, we know $P(\mathbf{S}_d) = 1 - (P(\mathbf{Z}_d) + P(\mathbf{D}_d))$.

We introduce some notation that we will need here and in the next section. Given a baseline level of influenza b, there is exactly one zero-outbreak model, which we denote by $\langle b \rangle$. A baseline level of influenza b, a start day s, and SEIR model parameters $\bar{\lambda}$ specify a one-outbreak model, which we denote by $\langle b, s, \bar{\lambda} \rangle$. A baseline level of influenza b, start days s and t, and SEIR model parameters $\bar{\lambda}$ and $\bar{\gamma}$ specify a two-outbreak model, which we denote by $\langle b, s, \bar{\lambda}, t, \bar{\gamma} \rangle$.

Now, we will derive $P(E(1:d)|\mathbf{Z}_d), P(E(1:d)|\mathbf{S}_d)$, and $P(E(1:d)|\mathbf{D}_d)$. We assume that the baseline level of influenza b is fixed. There is only one zero-outbreak model $\langle b \rangle$ so $P(E(1:d)|\mathbf{Z}_d) = P(E(1:d)|\langle b \rangle)$. For each start day d, we need to integrate over a continuum of one-outbreak outbreak models:

$$P(E(1:d)|\mathbf{S}_d) = \sum_{s=1}^{d}\left[\left(\frac{f(s)}{\sum_{i=1}^{d} f(i)}\right)\int_{\overline{\lambda}} P(E(1:d)|\langle b,s,\overline{\lambda}\rangle)p(\overline{\lambda})\,d\overline{\lambda}\right] \quad (14)$$

Likewise, we need to integrate over two-outbreak models to obtain:

$$P(E(1:d)|\mathbf{D}_d) = \sum_{s=1}^{d}\sum_{t=1}^{d}\left[\frac{f(s)f(t)}{\left(\sum_{i=1}^{d} f(i)\right)^2}\int_{\overline{\lambda\gamma}} P(E(1:d)|\langle b,s,\overline{\lambda},t,\overline{\gamma}\rangle)p(\lambda)p(\overline{\gamma})\,d\overline{\lambda}\overline{\gamma}\right] \quad (15)$$

In these equations, $p(\cdot)$ is a probability density function over SEIR model parameters. Note that we assume the double outbreak start dates are independent of one another, although this assumption can be relaxed.

If we assume that start days and SEIR model parameters are uniformly distributed over their ranges, we can estimate the integrals in Eqs. (14) and (15) as follows:

$$P(E(1:d)|\mathbf{S}_d) \approx \frac{1}{dN}\sum_{s=1}^{d}\sum_{n=1}^{N} P(E(1:d)|\langle b,s,\overline{\lambda}_n\rangle) \quad (16)$$

$$P(E(1:d)|\mathbf{D}_d) \approx \frac{1}{d^2N}\sum_{s=1}^{d}\sum_{t=1}^{d}\sum_{n=1}^{N} P(E(1:d)|\langle b,s,\overline{\lambda}_n,t,\overline{\gamma}_n\rangle) \quad (17)$$

where N is a large integer to determine enough points to estimate the integral, and the $\overline{\lambda}_n$ and $\overline{\gamma}_n$ are randomly and uniformly selected from their ranges.

These two equations perform model averaging. Consider Eq. (16) to estimate the probability of $E(1:d)$ given the assumption of one outbreak. For each possible start day, it randomly selects N single-outbreak models and computes the posterior probability of the evidence for each of those models and averages their predictions. We perform similar operations for zero and two outbreaks and combine the results using Eqs. (7)–(9) to find the probability of zero, one, or two outbreaks given $E(1:d)$.

2.3.2. Characterizing outbreaks

Given evidence $E(1:d)$, MODS predicts zero, one, or two outbreaks have started by day d depending on whether $P(\mathbf{Z}_d|E(1:d)), P(\mathbf{S}_d|E(1:d))$, or $P(\mathbf{D}_d|E(1:d))$ (derived in the previous section) is greatest.

We also want to know the expected number of influenza cases on day d' given $E(1:d)$. In general, $d' > d$ since we want to project into the future beyond the evidence we have already collected. (We can have $d' \leqslant d$, however, if we want to assess what has happened to date.) If $flu(d')$ is the number of influenza cases on day d', then:

$$\begin{aligned} E(flu(d')|E(1:d)) = {} & E(flu(d')|\mathbf{Z}_d, E(1:d))P(\mathbf{Z}_d|E(1:d)) \\ & + E(flu(d')|\mathbf{S}_d, E(1:d))P(\mathbf{S}_d|E(1:d)) \\ & + E(flu(d')|\mathbf{D}_d, E(1:d))P(\mathbf{D}_d|E(1:d)) \end{aligned} \quad (18)$$

That is, we can find the expected number of influenza cases predicted by the zero-outbreak model, one-outbreak models, and two-outbreak models separately and combine them in a weighted sum.

We already know how to compute $P(\mathbf{Z}_d|E(1:d)), P(\mathbf{S}_d|E(1:d))$, and $P(\mathbf{D}_d|E(1:d))$. The other terms can be expanded as follows:

$$E(flu(d')|\mathbf{Z}_d, E(1:d)) = \langle b\rangle(d') \quad (19)$$

$$E(flu(d')|\mathbf{S}_d, E(1:d)) = \sum_{s=1}^{d}\left[\frac{f(s)}{\sum_{j=1}^{d} f(j)}\int_{\overline{\lambda}} \langle b,s,\overline{\lambda}\rangle(d')p(\langle b,s,\overline{\lambda}\rangle|\mathbf{S}_d, E(1:d))\,d\overline{\lambda}\right] \quad (20)$$

$$E(flu(d')|\mathbf{D}_d, E(1:d)) = \sum_{s=1}^{d}\sum_{t=1}^{d}\left[\frac{f(s)f(t)}{\left(\sum_{j=1}^{d} f(j)\right)^2}\int_{\overline{\lambda\gamma}} \langle b,s,\overline{\lambda},t,\overline{\gamma}\rangle(d')p(\langle b,s,\overline{\lambda},t,\overline{\gamma}\rangle|\mathbf{D}_d, E(1:d))\,d\overline{\lambda\gamma}\right] \quad (21)$$

where $\langle b\rangle(d), \langle b,s,\overline{\lambda}\rangle(d)$ and $\langle b,s,\overline{\lambda},t,\overline{\gamma}\rangle(d)$ are the number of influenza cases predicted by each of the models on day d. Of course, $\langle b\rangle(d) = b$ is a constant function.

Finally, if we apply Bayes' rule, assume all start days are equally likely, assume the density functions are uniform, then we can approximate the integrals in Eqs. (20) and (21) with:

$$E(flu(d)|\mathbf{S}_d, E(1:d)) \approx \frac{1}{d}\sum_{s=1}^{d}\sum_{n=1}^{N}\left[\langle b,s,\overline{\lambda}_n\rangle(d)\frac{P(E(1:d)|\langle b,s,\overline{\lambda}_n\rangle)}{\sum_{m=1}^{N} P(E(1:d)|\langle b,s,\overline{\lambda}_m\rangle}\right] \quad (22)$$

$$E(flu(d)|\mathbf{D}_d, E(1:d)) \approx \frac{1}{d^2}\sum_{s=1}^{d}\sum_{t=1}^{d}\sum_{n=1}^{N}\left[\langle b,s,\overline{\lambda}_n,t,\overline{\gamma}_n\rangle(d)\frac{P(E(1:d)|\langle b,s,\overline{\lambda}_n,t,\overline{\gamma}_n\rangle)}{\sum_{m=1}^{N} P(E(1:d)|\langle b,s,\overline{\lambda}_m,t,\overline{\gamma}_m\rangle}\right] \quad (23)$$

where the $\overline{\lambda}_n\,\overline{\lambda}_m, \overline{\gamma}_n$, and $\overline{\gamma}_m$ are randomly and uniformly selected.

Again, these equations perform model averaging. Consider Eq. (22) to estimate the number of infectious on day d given $E(1:d)$ and the assumption of one outbreak. For each possible start day $s \leqslant d$ it randomly selects N single-outbreak models. For each of these models, it computes the number of infectious people predicted by the model times the probability of the model, then it computes the expected number of infectious on day d over these models that start on day s. The outer sum computes the expected number of infectious on day d over all the possible start days. We perform similar operations for zero and two outbreaks and combine the results using Eq. (18) to find the expected number of infectious.

2.4. Algorithmic methods

2.4.1. Prior probabilities and parameters

Population is the number of people in the region being monitored. S, E, and I are the initial values of *Susceptible*, *Exposed*, and *Infectious*. *Latent Period* and *Infectious Period* are expressed in terms of days. We set the prior probabilities for zero, one, or two outbreaks in the next year to be 0.1, 0.8, and 0.1, respectively. To generate a SEIR model of influenza we randomly and uniformly select parameters from Table 2. We also eliminate models with duration more than 180 days.

Table 2
SEIR parameter ranges.

Parameter	Minimum Value	Maximum Value
S	$0.5 * population$	$1.0 * population$
E	0	0
I	1	100
R_0	1.1	3.0
Latent Period	1	4
Infectious Period	1	8

2.4.2. Implementation

We now describe the overall operation of MODS. Given evidence $E(1:d)$ for days 1 through d, MODS creates the baseline model, 10,000 single-outbreak models with start day on or before day d, and 10,000 double-outbreak models with start days on or before day d. MODS then computes the probability of $E(1:d)$ given each of these models. It then computes the posterior probability of zero, one, or two outbreaks given $E(1:d)$. It uses the same models to compute the expected number of infectious on each day through the rest of the year. Note that models are *evaluated* using evidence from on or before day d to find their probabilities given $E(1:d)$. The models are then *projected* into the future and their predictions combined according to how well they performed on evidence in the past. The complexity and runtime of MODS are described in Appendix C.

3. Results

3.1. Experiments with real outbreaks

We applied MODS to clinical data from patient care reports in Allegheny County, Pennsylvania during the 2009–2010 influenza year and reports from Salt Lake City, Utah during the 2010–2011 influenza year. The ED data we monitor capture about 62% of patients in Allegheny County and 55% of patients in Salt Lake City. We estimated that approximately 1% of people infectious with influenza in the general population will visit an ED [4]. Baseline levels of influenza, NI-ILI, and other diseases were estimated from the summer months when there were no outbreaks.

We evaluated MODS' predictions both locally and regionally. Locally, we compared MODS' predictions to a seven day central moving average of the daily number of positive PCR tests in the monitored ED's. We will refer to the peak day of this moving average as the "local peak." Regionally, we compared MODS' predictions to the weekly totals of laboratory confirmed cases of influenza reported by the CDC's FluView system [19] for HHS Region 3 which includes Pennsylvania and HHS Region 8 which includes Utah.[1] We will refer to the Wednesday of a week with a maximum number of confirmed cases as the "regional peak." (We selected a specific day of the week in order to compare the local and regional peaks.)

Fig. 4 shows the probability of zero, one, or two outbreaks compared to the number of positive PCR tests performed in the monitored ED's, and the expected number of influenza cases, for Allegheny County from June 1, 2009 to March 31, 2010. Table 3 compares MODS' predictions to the local and regional peaks for Allegheny County during the same period. For instance, the row labeled "10/1/2009" means that on October 1 MODS predicted one influenza outbreak with a predicted peak three days after the local peak and one day after the regional peak. (There was a single local peak on October 18 and a single regional peak on October 20. These are essentially identical since the regional dates have only one week of precision.) Fig. 5 shows MODS' predictions on September 1, October 1, November 1, and December 1 with local PCR counts for comparison.

Fig. 6 shows the probability of zero, one, or two outbreaks compared to the number of positive PCR tests performed in the monitored ED's for Salt Lake City from June 1, 2010 to May 31, 2011. Table 4 compares MODS' predictions to the local and regional peaks for Salt Lake City during the same period. For instance, the row labeled "2/1/2011" means that on February 1 MODS predicted

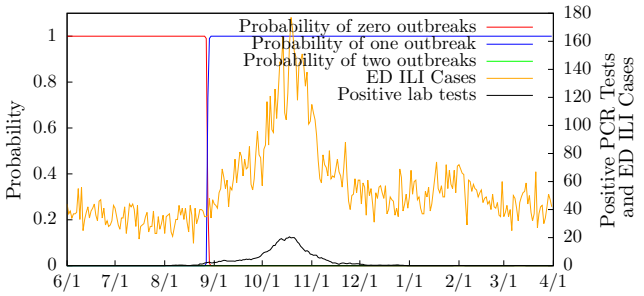

Fig. 4. Probability of zero, one, or two outbreaks for Allegheny County 2009–2010 influenza season.

Table 3
Predictions for Allegheny County and HHS Region 3 2009–2010 season.

Date	Predicted Number of Outbreaks	Local Peak Offset	Regional Peak Offset
8/1/2009	zero		
9/1/2009	one	−13	−15
10/1/2009	one	+3	+1
11/1/2009	one	+2	0
12/1/2009	one	+2	0
1/1/2010	one	+5	+3
2/1/2010	one	+4	+2
3/1/2010	one	0	−2
3/31/2010	one	+6	+4

two influenza outbreaks. The first predicted peak was two days before the first local peak and the second predicted peak seven days after the second local peak. Also, the first predicted peak was five days before the first regional peak and the second predicted peak four days after the second regional peak. (There were two local peaks on December 25, 2010 and February 19. 2011, and two regional peaks on December 28, 2010 and February 22, 2011. These are essentially identical since the regional dates have only one week of precision.) Fig. 7 shows how MODS' predictions evolve during the month of January when it first detects a second outbreak.

3.2. Experiments with synthetic outbreaks

Simulated outbreaks were generated by instantiating models of influenza outbreaks in the population with data from actual patient cases. To populate these simulated datasets we relied on the results of the polymerase chain reaction (PCR) test, a definitive indicator of influenza, to create three sets of patients:

- *Influenza* patients with a positive PCR test. Patients for this set were selected based on a positive PCR test result, but that test result was not included in the findings or computation of the likelihood.
- *NI-ILI* patients with a high probability of ILI but a negative PCR test. Again, patients for this set were selected based on the negative PCR test result, but that test result was not included in the findings or computation of the likelihood.
- *Other* patients from summer months with a low probability of ILI and no PCR test ordered.

Note that some patients in the *Influenza* set have a higher likelihood of NI-ILI and some patients in the *NI-ILI* set have a higher likelihood of influenza since the only definitive piece of information (the result of the PCR test) has been excluded. This reflects

[1] Region 3 is composed of Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, and West Virginia. Region 8 is composed of Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming.

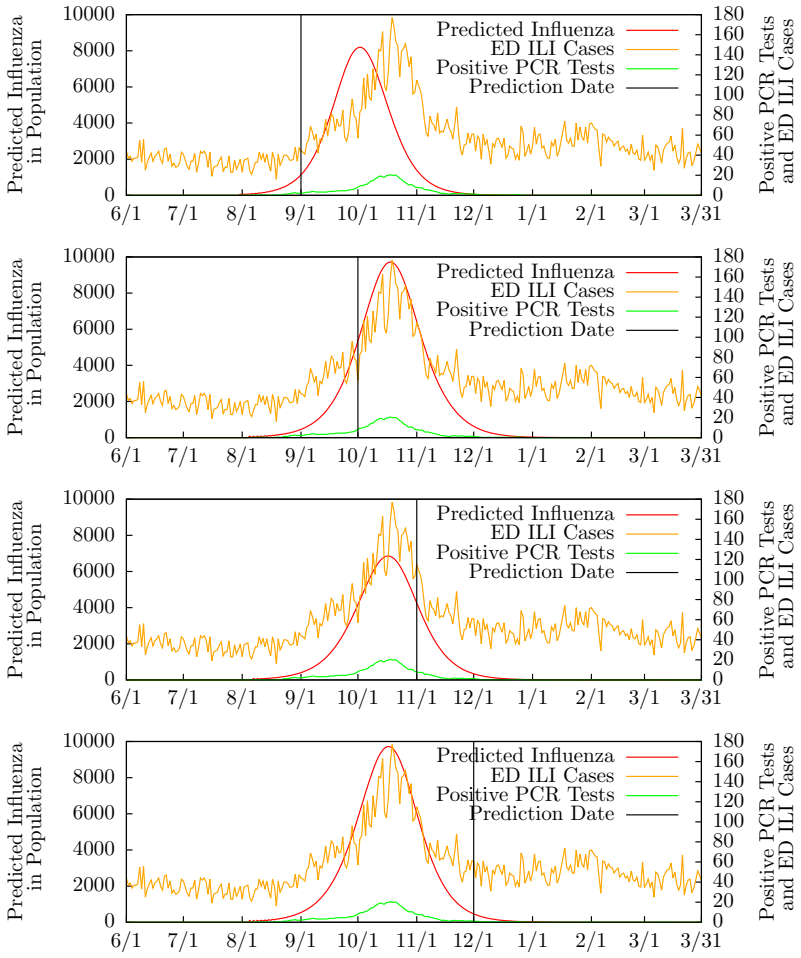

Fig. 5. Evolution of predictions for Allegheny County 2009–2010 influenza season.

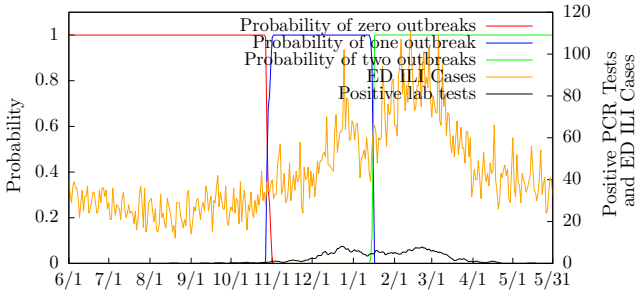

Fig. 6. Probability of zero, one, or two outbreaks for Salt Lake City 2010–2011 influenza season.

the fact that influenza and NI-ILI patients can have similar presentations.

We instantiate a model by assuming that each infectious case of influenza in the model goes to an ED with probability 0.01 (based on our estimate that 1% of the population with influenza will visit an ED). Thus, for each infectious case of influenza in the model we add a random element of *Influenza* to the test data with probability 0.01. Likewise, we assume that each day there are 10,000 NI-ILI cases and 50,000 other cases, each of whom go to a monitored ED with probability 0.01. Thus, for each of these cases we add a random element of *NI-ILI* or *Other* to the test data with probability 0.01. For instance, if a model predicts 1000 infectious influenza cases on day *d*, then the instantiation will contain approximately 10 *Influenza*, approximately 100 *NI-ILI*, and approximately 500 *Other* cases for day *d* as determined by binomial distributions.

Thus, these simulated outbreaks are *simulated* in the sense that the set of cases with influenza was determined by a model, but *real* in the sense that each patient case used real data from a confirmed case of influenza.

Table 5 summarizes the performance of MODS on 100 randomly generated single-outbreak datasets. We generated 100 single-outbreak models by randomly and uniformly selecting parameters from Table 2, instantiated each model as described above, and ran MODS on each one for each day from the start of the outbreak to its peak. The row starting with "0.100" represents the performance of MODS on the day when 0.100 of the total outbreak cases have occurred. This day will vary since the outbreaks were randomly generated, but the average number of days to the peak was 18, the average probability an outbreak was occurring was 1.00, and the mean error of estimating the peak was +10 (ten days *after* the actual peak).

To understand MODS' ability to recognize a second peak we conducted a set of experiments with two identical outbreaks with peaks that ranged from 5 to 65 days apart. Table 6 summarizes the results of these experiments. For instance, column 35 says that for a synthetic double outbreak with peaks 35 days apart, MODS began to predict with probability $\geqslant 0.50$ a second future peak 0.56 of the distance from the first peak to the second (about twenty days after the first peak and fifteen days before the second peak). MODS was unable to distinguish peaks only five days apart.

Table 4
Predictions for Salt Lake City 2010–2011 season.

Date	Predicted Number of Outbreaks	First Local Peak Offset	Second Local Peak Offset	First Regional Peak Offset	Second Regional Peak Offset
8/1/2010	zero				
9/1/2010	zero				
10/1/2010	zero				
11/1/2010	one	−17		−20	
12/1/2010	one	−8		−11	
1/1/2011	one	+6		+3	
1/10/2011	one	−2		−5	
1/20/2011	two	−6	−3	−9	−6
2/1/2011	two	−2	+7	−5	+4
3/1/2011	two	−4	+2	−7	−1
4/1/2011	two	+3	+12	0	+9
5/1/2011	two	+6	+1	+3	−2
5/31/2011	two	+6	+7	+3	+4

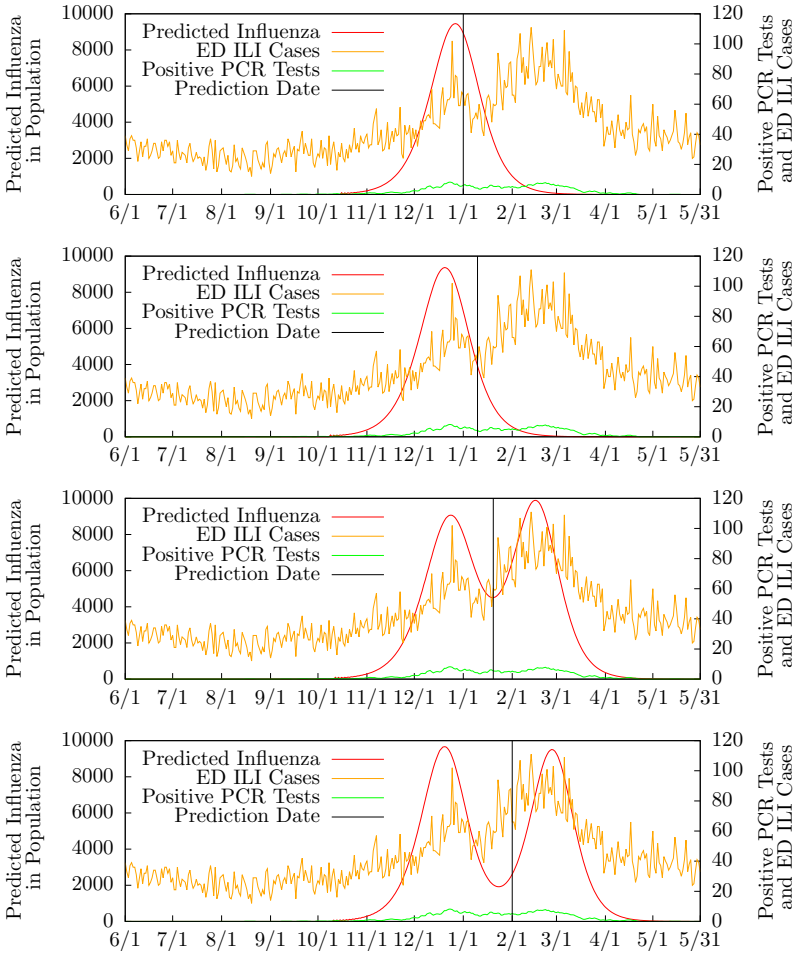

Fig. 7. Evolution of predictions for Salt Lake City 2010–2011 influenza season.

3.3. IRB approval

The research protocol was approved by both institutional IRBs (University of Pittsburgh PRO08030129 and Intermountain Healthcare 1024664). All the research patient data were de-identified.

4. Discussion

4.1. Experimental results

Tables 5 and 6 demonstrate how well MODS was able to predict actual single and double outbreaks. MODS' ability to predict single

Table 5
Results of evaluation on 100 simulated single outbreaks.

Fraction of Outbreak Cases That Have Occurred	Mean Number of Days Until Peak	Mean Posterior Probability an Outbreak is Occurring	Mean Error of Estimating Peak Day
0.001	55	0.09	−146
0.010	38	0.93	0
0.100	18	1.00	+10
0.250	9	1.00	+8
0.500	0	1.00	+6

Table 6
Discriminating one or two outbreaks.

Days Between Peaks	5	15	25	35	45	55	65
Tipping Point	–	1.00	0.66	0.56	0.56	0.54	0.55

influenza outbreaks is consistent with our prior work [4]. The double peak predicted by MODS from the Salt Lake City 2010–2011 data is intriguing. The *Utah Public Health Laboratory* tracks positive influenza laboratory tests broken down by type. Positive influenza A tests peaked around January 1 and positive influenza B tests peaked around February 26 [20]. These dates are close to the peaks predicted by MODS as shown in Fig. 8. This is noteworthy given that MODS did not have test information that distinguished A from B, but was able to recognize separate A and B outbreaks.

4.2. Limitations

The work reported here has several limitations. These stem from our basic assumptions, our use of SEIR models, and computational limitations.

We assume there is a constant, baseline number of influenza cases, b, in the general population throughout the year. We derive this number using data from summer months. However, the baseline varies daily and throughout the year. We could avoid this assumption by integrating over the range of values of b in Eqs. (14) and (15) each day. However, we do not believe this will affect the results since the baseline level is very small compared to outbreak levels.

We assume that the basic context of an outbreak remains fixed for its duration. However, in reality, weather and humidity, immunization programs, and shifting populations can affect how an outbreak proceeds [21]. These concerns can be addressed by allowing the basic parameters of models to change over time. We also assume that θ (the probability that a person in the general population who has influenza will go to an ED) is constant throughout the outbreak. In fact, it can change daily based on several factors, such as media reports. We can address this by integrating over a range of values for θ *each day* in Eq. (4) instead of just once over the entire outbreak period (essentially by moving the integral to after the daily product). However, we would want to model some dependency in θ from one day to the next.

We rely on the *mass action principle* that any two individuals in the population are equally likely to come in contact with each other. Although this is a common simplifying assumption, it ignores the fact that children are more likely to contact other children in the classroom and adults are more likely to contact adults at work. We can address this assumption by using stratified models [17] that maintain separate compartments for different demographic groups. More generally, since our framework only relies on specifying a model of influenza by a set of numeric parameters, we can explore different kinds of models (including stochastic models) that better capture the dynamics of actual outbreaks.

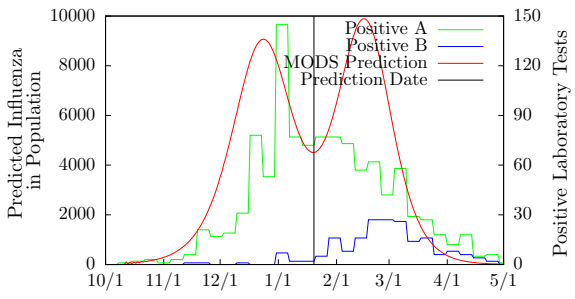

Fig. 8. Comparison of MODS' prediction on January 20 to positive influenza A and B laboratory tests in Salt Lake City 2010–2011. (Data from Utah's *Unified State Laboratory: Public Health*.)

Although the theory described above can be extended to any number of outbreaks, we have limited MODS to only two, primarily due to computational limitations. Although the two-outbreak restriction is sufficient to model the type A and type B waves of influenza that often occur, it ignores more complicated situations with multiple introductions of single or multiple strains. The size of the search space grows exponentially with the number of outbreaks considered. The simple Monte Carlo algorithm used here is unlikely to scale to the larger, more complex search space (although we might be able to handle three outbreaks by using a high-performance computer and switching to SIR models to eliminate some variables). Properly addressing this will require more sophisticated search algorithms.

4.3. Further work

The approach to outbreak detection and characterization described in this paper is flexible. We will briefly describe several possible extensions.

Incorporating different sources of data. Emergency department patient care reports are a rich source of information about outbreaks in the general population, but they are limited. Only a small fraction of people with influenza seek care in an emergency department, the hospitals we monitor are urban and suburban so many rural areas are invisible to us, and different demographic groups may seek ED care at different rates, skewing our data. Although no single source of data is without limitations or bias, we can correct for some of their individual shortcomings by combining them. The Bayesian framework described here provides a principled way to do this. Suppose that our evidence consists of two sets of data, E_1 and E_2. For instance, E_1 may be emergency department reports and E_2 may be daily counts of internet searches related to influenza [9]. If we assume that E_1 and E_2 are independent given the level of ILI in the population, we have $P(E_1E_2|\mathcal{I},\mathcal{N}) = P(E_1|\mathcal{I},\mathcal{N})P(E_2|\mathcal{I},\mathcal{N})$. The first term can be evaluated using Eq. (4). The second term can be evaluated similarly *if we know how many people with ILI will search for information about influenza on the internet each day*. We can get a first-order approximation to that quantity by relying on previous serological studies to determine the total number of people affected by influenza during one season [22] and internet search records to determine the number of influenza-related searches during an outbreak and non-outbreak periods.

Modeling influenza A and B separately. MODS first detected a second outbreak in the Salt Lake City 2010–2011 data on January 20. However, positive influenza B laboratory tests began to appear much earlier. MODS can be extended to recognize the second outbreak much sooner than it did. CDS can model influenza A and B as separate diseases. The presence of a subtyped laboratory test can distinguish A or B definitively and clinical presentation can also help distinguish them [23]. Given separate likelihoods for influenza A and B, MODS can model them as separate outbreaks. With these changes, we believe that MODS will be able to recognize the second outbreak much sooner than it did.

Efficient computational methods. The theory described here can be extended to any number of outbreaks. However, building a

practical system that can model complex multiple outbreaks will require more sophisticated algorithms.

Robust prediction of the effect of an intervention. MODS does not select a single most likely model and base its predictions on it. Rather, it entertains a set of plausible models, assigns a probability to each of them, then computes the expected trajectory of an outbreak based on a combination of the probabilities and predictions of all of the models. It is noncommittal. However, when public health officials are faced with an outbreak they often commit to a model they believe best describes the outbreak and design an intervention based on it [11,12,24,25]. This is fraught with danger since they must make a decision with huge public health, economic, and political risk based on limited evidence early in an outbreak [2]. We propose a different approach: instead of picking a single outbreak model, use *all* of the models, project the effect of an intervention against each model in the set, and predict the interventions' efficacy with the expected outcome based on the models' probabilities that were computed before the intervention.

5. Conclusions

We described a framework for the detection and characterization of multiple, overlapping outbreaks of influenza from ED patient care reports. We also described the implementation and evaluation of a system capable of detecting and characterizing two overlapping outbreaks. We believe that the results reported here show that the approach we describe is sufficient to detect and characterize overlapping outbreaks. Further work will includes investigating more sophisticated models of multiple outbreaks, data that can distinguish separate outbreaks, and more efficient computational methods.

Acknowledgements

This work was supported by NIH grant R01 LM011370 on *Probabilistic Disease Surveillance*. John Aronis was supported by the National Library of Medicine Training Grant T15 LM007059-31 to the University of Pittsburgh. Part of this work was done on the Olympus High Performance Compute Cluster located at the Pittsburgh Supercomputing Center at Carnegie Mellon University, which is supported by National Institute of General Medical Sciences Modeling Infectious Disease Agent Study (MIDAS) Informatics Services Group grant 1U24GM110707.

Conflict of interest

Authors declare that there is no conflict of interest.

Appendix A. CDS

The data consist of the full text of emergency department patient care reports collected at two sites: the University of Pittsburgh Medical Center in Allegheny County (AC), Pennsylvania and Intermountain Healthcare in Salt Lake City (SLC), Utah. There are approximately 300–600 patient care reports each day at each site. A patient care report may consist of the union of several separate documents filed during the patient's stay in the ED or shortly thereafter. We ignore administrative delays and assume the entire set of documents is available when the patient is discharged from the ED.

Each report is processed by a site-specific parser to extract a set of 79 findings that were identified by expert clinicians as potentially relevant to influenza like illness. The parsers are implemented using Topaz 2.0 [5] and apply pattern-matching and deduction rules to extract clinical findings. The input to each parser is a patient care report. The output is a set of clinical findings. Findings that are not found and extracted from the text are given the value "M" for missing. After the text of the report has been processed, the patient's age and three laboratory results are added to the set of findings, for a total of 83 findings (see Table 1).

Two site-specific case detection systems were created for AC and SLC using the following procedure at each site [26]. Training datasets were created at each site using ED encounters between January 1, 2008 and May 31, 2010. Cases with a positive influenza lab test were labeled *flu*, cases with a negative influenza lab test were labeled *NI-ILI*, and other cases were labeled *other*. Features were selected based on information gain. Then, a Bayesian network (CDS) was constructed for each site using the K2 algorithm [27]. Thus, each CDS is a Bayesian network that represents the conditional probability distribution of the findings and disease state (influenza, NI-ILI, or other). Given the findings of a particular patient, $E(p,d)$, the appropriate CDS (AC or SLC) can produce likelihoods $P(E(p,d)|flu)$, $P(E(p,d)|NI\text{-}ILI)$, and $P(E(p,d)|other)$. That is, the likelihood of that patient's findings given they have influenza, non-influenza influenza-like illness (NI-ILI), or some other condition (such as appendicitis and trauma). MODS uses these likelihoods to evaluate models.

Table 1 shows the 79 features parsed from the text of patient care reports, plus age and the added laboratory variables. The annotation "AC" means that the feature was used in the CDS for Allegheny County, and the annotation "SLC" means that the feature was used in the CDS for Salt Lake City.

Appendix B. Modeling NI-ILI

We model NI-ILI with a Bayesian time-series. On the first day, we start with a prior probability distribution over the fraction of ED patients with NI-ILI $P(NI-ILI_1 = x)$ for $x = 0.00, 0.01, 0.02, \ldots, 1.00$. (For instance, $P(NI-ILI_1 = 0.20)$ is the probability that 20% of the ED patients on day 1 have NI-ILI.) Each day, d, we update our probability distribution for NI-ILI with the available lab results using the following equation:

$$P(NI-ILI_d = x|\oplus_d, \ominus_d) = \frac{P(\oplus_d, \ominus_d|NI-ILI_d = x)P(NI-ILI_d = x)}{\sum_{y=0.00}^{1.00} P(\oplus_d, \ominus_d|NI-ILI_d = y)P(NI-ILI_d = y)} \quad (B.1)$$

where $\oplus_d$ and $\ominus_d$ are the number of positive and negative tests on day d. It is likely that clinicians are more or less likely to order a lab test for a given patient as an outbreak progresses, so we simply assume that the relative rate of lab tests given influenza or NI-ILI, $P(lab\ ordered|flu)/P(lab\ ordered|NI\text{-}ILI)$, is a constant c. Estimating the number of influenza patients in the ED with $\theta \cdot \mathcal{I}(d)$, we have (by the binomial theorem):

$$P(\oplus_d, \ominus_d|NI-ILI_d = x) = \binom{\oplus_d + \ominus_d}{\ominus_d} \left(\frac{x|pts(d)|}{x|pts(d)| + c\theta \cdot \mathcal{I}(d)}\right)^{\ominus_d} \left(\frac{c\theta \cdot \mathcal{I}(d)}{x|pts(d)| + c\theta \cdot \mathcal{I}(d)}\right)^{\oplus_d} \quad (B.2)$$

That is, $P(\oplus_d, \ominus_d|NI-ILI_d = x)$ is the probability that $\oplus_d$ of the lab tests 'hit' influenza patients and $\ominus_d$ of the lab tests 'hit' NI-ILI patients. Then, each day we update our beliefs by $P(NI-ILI_{d+1} = x) = P(NI-ILI_d = x|\oplus_d, \ominus_d)$. Finally, we define $\mathcal{N}(d)$ to be the expected number of NI-ILI patients in the ED on day d:

$$\mathcal{N}(d) \approx |pts(d)| \sum_{x \in \{0.0, 0.1, \ldots, 1.0\}} xP(NI-ILI_d = x) \quad (B.3)$$

Note that $\mathcal{N}$ is computed day-by-day, updating our beliefs each day using only laboratory tests from the previous day and earlier, so $\mathcal{N}(d)$ depends only on $E(1:d-1)$.

Given data $E(1:d)$, we need $P(E(1:d)|\mathcal{I},\theta)$ to evaluate a model $\mathcal{I}$, where θ is the fraction of infectious influenza cases we expect to go to an ED on any given day. We can expand this as follows:

$$P(E(1:d)|\mathcal{I},\theta) = P(E(d)|E(1:d-1),\mathcal{I},\theta) \quad \text{(B.4)}$$
$$\times P(E(d-1)|E(1:d-2),\mathcal{I},\theta)$$
$$\cdots \quad \text{(B.5)}$$
$$\times P(E(1)|\mathcal{I},\theta)$$

However, each day d $E(d)$ does not depend on all of the evidence from previous days or the entire model $\mathcal{I}$, but only $\mathcal{N}(d)$ and $\mathcal{I}(d)$, so we have:

$$P(E(1:d)|\mathcal{I},\theta) = \prod_{d'=1}^{d} P(E(d')|\mathcal{I}(d'),\mathcal{N}(d'),\theta) \quad \text{(B.6)}$$

Appendix C. Complexity and runtime

The overall operation of MODS is to generate a model of influenza, create its corresponding model of NI-ILI, and evaluate the pair against the data. Then, repeat this for several thousand models. After the models have been created and evaluated, the expected number of infectious cases each day is computed. To generate a model of influenza, MODS randomly and uniformly selects parameters from the ranges in Table 2. A model can then be generated in $O(D)$ time where D is the number of days in the dataset being analyzed. Generating NI-ILI models requires a preprocessing step to count the number of positive and negative lab tests each day. This is done once in time $O(DP)$, where P is the average number of patients per day. For each model of influenza, creating its corresponding NI-ILI model requires us to evaluate Eq. (B.1) for each day and each value of n. This can be done in time $O(Dc_nc_t)$ where c_n is the number of values of n and c_t is the average number of lab tests administered each day. Finally, we compute $\mathcal{N}(d)$ from Eq. (B.3) in time $O(D)$. Thus, generating M influenza/NI-ILI models takes time $O(DP + M(D + Dc_nc_t))$. We can evaluate an influenza/NI-ILI model in time $O(c_\theta DP)$, where c_θ is the number of discrete values of θ and P is the average number of patients per day, by implementing Eq. (6) in the obvious way. We speed this up by precomputing and caching values of $P(E(d)|i,n)$ where i and n are discretized fractions of ED patients with influenza or NI-ILI. If c_d is the number of discrete values we can precompute the cache in time $O(DPc_d^2)$ then evaluate each model in time $O(D)$. Thus, scoring M models takes $O(DPc_d^2 + MD)$ time. After all the models have been created and evaluated we update the expected number of influenza cases in time $O(MD)$ for M models. Putting all of this together, a single run of MODS takes $O(DP + M(D + Dc_nc_t) + DPc_d^2 + MD) = O(MDc_nc_t + DPc_d^2)$ Typically, D is 365, P is about 500, c_d is 400, M is about 100,000, c_θ is three to five, c_n is 100, c_t is about 100. Creating the caches for the Allegheny County and Salt Lake City takes about six hours for each dataset, and running MODS takes about four hours on each dataset, when run on a single 1.6 GHz processor. The synthetic experiments were run on a high-performance cluster of 64 2.3 GHz processors with shared memory. It took about twelve hours to create and analyze 100 simulated single outbreaks, and about five hours to create and analyze 60 simulated double outbreaks.

References

[1] World Health Organization, Influenza Fact Sheet, 2003.
[2] Marc Lipsitch, Steven Riley, Simon Cauchemez, Azra C. Ghani, Neil M. Ferguson, Managing and reducing uncertainty in an emerging influenza pandemic, N. Engl. J. Med. 361 (2) (2009) 112–115.
[3] Michael Wagner, Fuchiang Tsui, Gregory Cooper, Jeremy Espino, Hendrik Harkema, John Levander, Ricardo Villamarin, Ronald Voorhees, Nicholas Millett, Christopher Keane, Probabilistic, decision-theoretic disease surveillance and control, Online J. Public Health Inform. 3 (3) (2011).
[4] Gregory F. Cooper, Ricardo Villamarin, Fu-Chiang Rich Tsui, Nicholas Millett, Jeremy U. Espino, Michael M. Wagner, A method for detecting and characterizing outbreaks of infectious disease from clinical reports, J. Biomed. Inform. 53 (2015) 15–26.
[5] Fuchiang Tsui, Michael Wagner, Gregory Cooper, Jialan Que, Hendrik Harkema, John Dowling, Thomsun Sriburadej, Qi Li, Jeremy Espino, Ronald Voorhees, Probabilistic case detection for disease surveillance using data in electronic medical records, Online J. Public Health Inform. 3 (3) (2011).
[6] Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, Antoine Flahault, Prediction of the spread of influenza epidemics by the method of analogues, Am. J. Epidemiol. 158 (10) (2003) 996–1006.
[7] Jimmy Boon Som Ong, I. Mark, Cheng Chen, Alex R. Cook, Huey Chyi Lee, Vernon J. Lee, Raymond Tzer Pin Lin, Paul Ananth Tambyah, Lee Gan Goh, Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore, PloS One 5 (4) (2010).
[8] Jeffrey Shaman, Alicia Karspeck, Forecasting seasonal outbreaks of influenza, Proc. Nat. Acad. Sci. 109 (50) (2012) 20425–20430.
[9] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, Larry Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (7232) (2009) 1012–1014.
[10] I.M. Hall, R. Gani, H.E. Hughes, S. Leach, Real-time epidemic forecasting for pandemic influenza, Epidemiol. Infect. 135 (3) (2007) 372–385.
[11] David Greenhalgh, Control of an epidemic spreading in a heterogeneously mixing population, Math. Biosci. 80 (1) (1986) 23–45.
[12] Andreas Handel, Ira M. Longini, Rustom Antia, What is the best control strategy for multiple infectious disease outbreaks?, Proc Roy. Soc. Lond. B: Biol. Sci. 274 (1611) (2007) 833–837.
[13] Christina E. Mills, James M. Robins, Carl T. Bergstrom, Marc Lipsitch, Pandemic influenza: risk of multiple introductions and the need to prepare for them, PLoS Med. 3 (6) (2006).
[14] Olivia Prosper, Omar Saucedo, Doria Thompson, Griselle Torres-Garcia, Xiaohong Wang, Carlos Castillo-Chavez, Modeling control strategies for concurrent epidemics of seasonal and pandemic H1N1 influenza, Math. Biosci. Eng. 8 (1) (2011) 141–170.
[15] Centers for Disease Control and Prevention, Considerations for distinguishing influenza-like illness from inhalational anthrax, 2001.
[16] P.L. Elkin, D.A. Froehling, D.L. Wahner-Roedler, S.H. Brown, K.R. Bailey, Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes, Ann. Intern. Med. 156 (2012).
[17] Emilia Vynnycky, Richard White, An Introduction to Infectious Disease Modelling, Oxford University Press, 2010.
[18] Centers for Disease Control and Prevention, Flu activity and surveillance, 2016.
[19] Centers for Disease Control and Prevention, FluView Interactive, 2016.
[20] Utah Department of Health, Utah MMWR Week 20 of 2011, 2011.
[21] Jeffrey Shaman, Virginia E. Pitzer, Cécile Viboud, Bryan T. Grenfell, Marc Lipsitch, Absolute humidity and the seasonal onset of influenza in the continental United States, PLoS Biol. 8 (2) (2010).
[22] Carrie Reed, Jacqueline M. Katz, Kathy Hancock, Amanda Balish, Alicia M. Fry, Prevalence of seropositivity to pandemic influenza A/H1N1 virus in the United States following the 2009 pandemic, PloS One 7 (10) (2012).
[23] Masahide Kaji, Aya Watanabe, Hisamichi Aizawa, Differences in clinical features between influenza A H1N1, A H3N2, and B in adult patients, Respirology 8 (2) (2003) 231–233.
[24] Raymond Gani, Helen Hughes, Douglas Fleming, Thomas Griffin, Jolyon Medlock, Steve Leach, Potential impact of antiviral drug use during influenza pandemic, Emerg. Infect. Dis. 11 (9) (2005) 1355–1362.
[25] Ira M. Longini, Azhar Nizam, Shufu Xu, Kumnuan Ungchusak, Wanna Hanshaoworakul, Derek A.T. Cummings, M. Elizabeth Halloran, Containing pandemic influenza at the source, Science 309 (5737) (2005) 1083–1087.
[26] Ye Ye, Michael M. Wagner, Gregory F. Cooper, Jeffrey P. Ferraro, H. Su, P. Gesteland, Peter J. Haug, Nicholas E. Millett, John M. Aronis, Andrew J. Nowalk, Victor M. Ruiz, Arturo L. Pineda, Lingyun Shi, Rudy Van Bree, A study of the transferability of influenza case detection systems between two large healthcare systems, PLoS One 12 (4) (2017).
[27] Gregory Cooper, Edward Herskovits, A Bayesian method for the induction of probabilistic networks from data, Mach. Learn. 9 (1992).