Original Paper

A Bayesian System to Detect and Track Outbreaks of Influenza-Like Illnesses Including Novel Diseases: Algorithm Development and Validation

John M Aronis[1], PhD; Ye Ye[1], MSPH, PhD; Jessi Espino[1], MS, MD; Harry Hochheiser[1], PhD; Marian G Michaels[2], MPH, MD; Gregory F Cooper[1], MD, PhD

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

[2]Department of Pediatrics, University of Pittsburgh School of Medicine, UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:
John M Aronis, PhD
Department of Biomedical Informatics
University of Pittsburgh
5607 Baum Blvd, Suite 500
Pittsburgh, PA, 15206-3701
United States
Phone: 1 412 624 5100
Email: jma18@pitt.edu

Abstract

Background: The early identification of outbreaks of both known and novel influenza-like illnesses (ILIs) is an important public health problem.

Objective: This study aimed to describe the design and testing of a tool that detects and tracks outbreaks of both known and novel ILIs, such as the SARS-CoV-2 worldwide pandemic, accurately and early.

Methods: This paper describes the ILI Tracker algorithm that first models the daily occurrence of a set of known ILIs in hospital emergency departments in a monitored region using findings extracted from patient care reports using natural language processing. We then show how the algorithm can be extended to detect and track the presence of an unmodeled disease that may represent a novel disease outbreak.

Results: We include results based on modeling diseases like influenza, respiratory syncytial virus, human metapneumovirus, and parainfluenza for 5 emergency departments in Allegheny County, Pennsylvania, from June 1, 2014, to May 31, 2015. We also include the results of detecting the outbreak of an unmodeled disease, which in retrospect was very likely an outbreak of the enterovirus D68 (EV-D68).

Conclusions: The results reported in this paper provide support that ILI Tracker was able to track well the incidence of 4 modeled influenza-like diseases over a 1-year period, relative to laboratory-confirmed cases, and it was computationally efficient in doing so. The system was also able to detect a likely novel outbreak of EV-D68 early in an outbreak that occurred in Allegheny County in 2014 as well as clinically characterize that outbreak disease accurately.

(JMIR Public Health Surveill 2024;10:e57349) doi: 10.2196/57349

KEYWORDS

biosurveillance; outbreak; novel disease; natural language processing; disease modeling; Bayesian modeling; influenza; influenza-like illnesses; novel diseases; public health; COVID-19; SARS-CoV-2; coronavirus; hospital; hospitals; emergency department; patient care; NLP; algorithm; respiratory syncytial; human metapneumovirus; parainfluenza; Pennsylvania; enterovirus D68; surveillance

Introduction

Background

Respiratory viruses are responsible for annual and oftentimes overlapping outbreaks in human populations. This overlapping disease activity confounds the diagnosis and treatment of patients presenting with influenza-like illnesses (ILIs) and the associated high caseloads stress the clinical and logistical capacity of the health care system. Thus, accurately detecting and tracking overlapping outbreaks due to these viruses are important tasks with public health implications and clinical repercussions for those at high risk [1-5]. The ideal surveillance system will notice an outbreak after just a few cases that may be distributed across several shifts at multiple hospitals. An individual physician may see just 1 or 2 cases, which might seem inconsequential to them and not worthy of mention, but an automated surveillance system, such as we describe here, can gain statistical power and timeliness by aggregating data across an entire region. The proposed system can harness the entire set of patients and their symptoms who present to all the emergency departments (EDs) in a region. By harnessing the sheer volume of this information, it may recognize cases and patterns that would elude human observers early in the outbreak, and thus, serve as a sentinel to detect and characterize outbreaks early.

In addition to detecting and tracking known viruses of concern, the world is also faced with the emergence of novel viruses (or the re-emergence of previously quiescent viruses) and the diseases that they cause, as evidenced by the appearance of the SARS-CoV-2 worldwide pandemic. Early detection and tracking of a novel or re-emerging outbreak disease can be critical in informing both the care of individual patients and the decisions made by public health officials. While we hope to someday prevent the emergence of pathological viruses before they strike the human population [6], a more realistic goal for the near term is early detection and tracking [7]. Modeling known diseases that we expect to see provides a useful background against which to detect the emergence of new diseases that have a different clinical or epidemiological presentation.

Previous Work

The most promising route to the early detection of outbreaks of pathological viruses is real-time surveillance of human and animal populations [2]. There are myriad approaches to the problem of disease surveillance based on different data sources and technologies. Most recent work has been based on data from laboratories [8] and sentinel physicians [9]. Social media, including Google Flu Trends, have been proposed as a useful tool [10], but initial efforts have not worked as well as expected [11-14]. More recent approaches to the use of social media are promising [15-18]. A variety of other sources, such as sales of over-the-counter medications, absenteeism, and traffic patterns have also been proposed [19].

Most of these data sources have significant limitations. For instance, laboratory tests of infectious pathogens can identify outbreak diseases that are known and tested for, but such tests are blind to emerging pathogens. Furthermore, laboratory tests are not always routinely performed or reported, especially in resource-challenged health care settings. Sentinel physicians and various “drop-in” surveillance methods have poor coverage, and individual physicians might not notice or appreciate isolated cases [20]. Methods based on data from social media [21], absenteeism [22], traffic patterns [23], etc, are nonspecific [24]. More importantly, these methods will only recognize anomalies after a substantial number of people have been affected, losing important time when an infectious disease could have been identified.

Syndromic surveillance [25] seeks to identify clusters of signs and symptoms among patients recorded during routine medical care, especially in hospital EDs. This approach has the advantages of both broad coverage (nearly every community in the United States is served by some ED) and the use of clinical information (including chief complaints, vital signs, clinical findings, etc). Unlike traditional systems that rely on voluntary reports from health care providers (who work with individual cases), syndromic surveillance systems use data about entire populations that are continuously and automatically acquired. Syndromic surveillance attains much of its timelines from the identification of syndromes before confirmed diagnoses are made.

To detect outbreaks of rare or novel threats, however, we must go beyond the traditional syndromic surveillance systems that only detect known syndromes, such as the Centers for Disease Control and Prevention (CDC) National Syndromic Surveillance Program [26]. To identify novel threats to public health, syndromic surveillance systems need to identify clusters of patients even if they are not characterized by a predefined syndromic grouping. The North Carolina Disease Event Tracking and Epidemiologic Collection Tool system [27] identifies clusters of related patients based on time of arrival to identify clusters of related ED visits in 30- and 60-minute windows. The study by Burkom et al [28] used a Fisher exact test to identify anomalous clinical terms in an 8-hour block of current chief complaints compared with a 30-day sliding baseline. Sets of anomalous terms are then presented to a human monitor for further investigation. The multidimensional semantic scan system [29] uses latent Dirichlet allocation to learn a set of syndromes directly from ED chief complaints. The learned syndromes include 25 “static” topics that correspond to common health conditions and a set of 25 “emerging” topics from recent data that may indicate newly emerging threats. The multidimensional semantic scan system also uses practitioner feedback to distinguish between relevant and irrelevant clusters. The system was extensively tested on data from New York City. In previous work [30], we described a system that builds a probabilistic model of normal (baseline) ILI activity using a large set of patient findings extracted from patient care reports with natural language processing. It then looks for statistically significant deviations from baseline normal activity. This system does not rely on just a small set of findings as might be extracted from patients’ chief complaints.

The ILI Tracker algorithm introduced here differs from previous work in several important ways. First, it uses a large set of findings extracted from patient care reports, not just chief complaints. This is important because most ILIs cannot be distinguished solely based on chief complaints but require a

complete assessment to be recognized. This is especially important for the recognition of novel or emergent diseases that may only be detected by the presence of uncommon or unusual combinations of findings. Second, ILI Tracker explicitly models and tracks known ILIs. This provides a backdrop against which novel diseases can be detected even in the presence of modeled ILIs. Finally, because the ILI Tracker explicitly models known diseases and their symptoms, it can identify and characterize patients who do not fit the profile of the expected modeled ILIs and bring them to the attention of clinicians for further evaluation.

Current Work

This paper describes the ILI Tracker algorithm that tracks the daily occurrence of a set of modeled ILIs in hospital EDs in a monitored region using natural language processing on patient care reports. A set of clinical findings is extracted from full-text patient care reports that are available at the time of care or shortly thereafter. These findings are used by machine learning algorithms to learn probabilistic models of a set of diseases. The models are used to determine the likelihood of each disease for each patient [31]. These likelihoods are then used to compute the expected prevalence of each disease in the EDs on each day. ILI Tracker also analyzes whether recent patient cases in the EDs are not well explained by the known diseases that it models. If so, it suggests the possible presence of a novel outbreak of a disease in the population.

The remainder of this paper first describes the ILI Tracker algorithm in detail. We build models for diseases such as influenza, respiratory syncytial virus (RSV), human metapneumovirus (hMPV), and parainfluenza (PIV) using data from 5 EDs in Allegheny County, Pennsylvania, from June 1, 2010, to May 31, 2014. We then present initial experiments on how well it can track those diseases from June 1, 2014, to May 31, 2015. Finally, we present a preliminary investigation of an algorithm based on the ILI Tracker for detecting the presence of an unmodeled disease.

Methods

The Algorithm

To diagnose a patient, a clinician must consider that patient's findings as well as the prevalence of various diseases in the community. For instance, given a patient with a fever and cough, a high rate of influenza in the population will elevate the probability that the patient has influenza, whereas a high rate of SARS-CoV-2 will elevate the probability of SARS-CoV-2. The situation becomes more complicated when multiple viruses (with similar or overlapping symptoms) are circulating in the environment for which the rate of each must be accounted. Bayesian inference provides a principled way to do this.

We assume each patient has exactly one of the diseases {dx_0, dx_1, …, dx_n}. If $Pr(dx_i \mid findings)$ is the probability that a patient has disease dx_i (for some i in $0,\ldots,n$) given their findings, and $Pr(dx_i)$ is the prevalence of that disease in the population at that time, then:

$$Pr(dx_i \mid findings) = \frac{Pr(findings \mid dx_i)Pr(dx_i)}{\sum_{j=0}^{n} Pr(findings \mid dx_j)Pr(dx_j)} \quad (1)$$

The denominator of Equation 1 is a sum over all the diseases we are modeling and is a normalizing factor, so we have:

$$P(dx_i \mid findings) \propto Pr(findings \mid dx_i)Pr(dx_i) \quad (2)$$

That is, given a patient's findings, the probability that the patient has disease dx_i is proportional to the product of the likelihood of their findings given dx_i times the prior probability of dx_i.

Using this formulation, we can compute the probability of each disease for each patient. The expected number of patients with each disease is the sum over the probability each patient has that disease. For example, suppose there are 50 patients and 20 have a 0.1 probability of influenza, while 30 have a 0.2 probability. Then the expected number of patients with influenza is 20×0.1+30×0.2=8. Given the expected number with each disease, we can compute the expected proportion of each disease. In the example above, the expected proportion of patients with influenza would be 8/50=0.16.

The ILI Tracker algorithm combines the above steps to compute the expected proportion of each disease each day. It starts with prior probabilities for each disease, computes the expected number of patients with each disease as above, and then uses the proportion of each disease as the prior probability of each disease on the next day. This process continues day by day. The remainder of this section provides the technical details of this algorithm.

We first introduce some notation. Let *days* be the sequence of days under consideration, *pts*(*d*) be the number of patients who visited the EDs on day *d*, *D*(*p*,*d*) be the set of findings ("data") for patient *p* on day *d*, and $D(d) = \{D(p,d)\}_{p=1}^{pts(d)}$ be all of the data for the patients on day *d*. Note that we number days starting with zero. We assume there is a set of modeled diseases $Dx=\{dx_0, dx_1, \ldots, dx_n\}$ where $\{dx_1, \ldots, dx_n\}$ are the diseases of interest and dx_0 denotes other known diseases. Here, other represents a large set of known diseases including trauma, cardiac events, diabetic emergencies, etc, that can occur and are not one of the *n* modeled diseases.

For a patient *p* on day *d* with findings *D*(*p*,*d*), we can calculate the probability they have a particular disease dx_i:

$$Pr(dx_i \mid D(p,d)) = \frac{Pr(D(p,d) \mid dx_i)Pr_d(dx_i)}{\sum_{j=0}^{n} Pr(D(p,d) \mid dx_j)Pr_d(dx_j)} \quad (3)$$

where $Pr_d(dx_i)$ is the prior probability of dx_i on day *d* and $Pr(D(p,d)|dx_i)$ is the likelihood of the patient *p*'s findings given they have disease dx_i. We describe how we compute each of these quantities below. We compute the expected number of patients with each dx_i on day *d* as:

$$E_d(dx_i) = \sum_{p=1}^{pts(d)} Pr(dx_i \mid D(p,d)) \quad (4)$$

We can now estimate the posterior probability of each disease on day d:

$$Pr_d(dx_i \mid D(d)) = \frac{E_d(dx_i) + m \times prior(dx_i)}{pts(d) + m} \quad (5)$$

where m is the so-called equivalent sample size and $prior(dx_i)$ is the prior probability of disease dx_i. The terms m and $prior(dx_i)$ in Equation 5 provide smoothing of the estimate that avoids relying too heavily on small values of $E_d(dx_i)$ and $pts(d)$ by augmenting the data with an additional m patients with diseases distributed according to $prior(dx_i)$. We specify m and $prior(dx_i)$ below. We then make the disease priors for day d+1 equal to the disease posteriors for day d.

In summary, the overall procedure is as follows. Set each prior probability $Pr_0(dx_i)$ to initial values as described below. Then for each day d:

1. Compute $Pr(dx_i \mid D(p,d))$ for each $p \in Pts(d)$ using Equation 3.

2. Compute $E_d(dx_i)$ for each dx_i using Equation 4.

3. Compute $Pr_d(dx_i \mid D(d))$ for each dx_i using Equation 5.

4. Set $Pr_{d+1}(dx_i) = Pr_d(dx_i \mid D(d))$ for each dx_i.

Steps 1-4 are repeated for each successive day. That is, each day, d (beginning with the first day d=0), we start with a prior probability, $Pr_d(dx_i)$, for each disease. We then use data from the patients in the EDs on day d to compute a posterior (updated) probability, $Pr_d(dx_i \mid D(d))$, for each disease. The posterior probability for each disease is then used as the prior probability for that disease, $Pr_{d+1}(dx_i)$, the next day.

Our data spans the time from June 1, 2010, to May 31, 2015. We use the data from June 1, 2010, to May 31, 2014, to build and train disease models, and we start monitoring on June 1, 2014. The prior probabilities on June 1, 2014 for influenza, RSV, hMPV, and PIV were set to 0.033, 0.035, 0.005, and 0.003, respectively. We determined these values by running ILI Tracker on the data for March 1, 2014 to May 31, 2014 with priors on March 1, 2014 of 0.05, 0.05, 0.05, and 0.05, and using the resulting posterior probabilities from May 31, 2014 as the prior probabilities for June 1, 2014. The equivalent sample size, m, was set to 10.

Detecting the Presence of Unmodeled Diseases

As mentioned, the better we can model the usual diseases we expect to see in the ED, the better we anticipate detecting novel diseases. The remainder of this section specifies how we do so.

We can regard the output of the ILI Tracker as a model of the types of patients who are in the ED each day. ILI Tracker assumes the presence of a fixed set of diseases that can be modeled using Bayesian networks with specified findings. If this assumption is satisfied then the model should explain the evidence (the patients and their findings) well and the probability of the data given by the model produced by ILI Tracker will be relatively high.

If any of these assumptions are violated, in particular, if there are patients with a novel, unmodeled disease, the probability of the data given by the model produced by ILI Tracker will likely be reduced compared with a previous period of time when only modeled diseases were present in the ED. If we track the probability of the data given by the output of ILI Tracker, a decrease in this daily probability may signal the presence of an unmodeled disease. An unmodeled disease may be a novel disease or a re-emergent disease that we are not currently modeling.

The day-to-day probabilities for each disease computed by ILI Tracker can be used to perform a posterior predictive check by computing the likelihood of the data each day given by the output of ILI Tracker as follows:

$$\Pr\left(D(d) \mid \{D(p,d)\}_{p=1}^{pts(d)}, \{\Pr_d(dx_i)\}_{i=0}^{n}\right) = \prod_{p=1}^{pts(d)} \sum_{i=0}^{n} \Pr(D(p,d) \mid dx_i) \Pr_d(dx_i) \quad (6)$$

where $\{D(p,d)\}_{p=1}^{pts(d)}$ are the findings on day d for each patient p and $\{D(p,d)\}_{p=1}^{pts(d)}$ is the set of prior probabilities of each disease on day d computed by ILI Tracker.

Let the null hypothesis be that the likelihood given by Equation 6 for the current day is the same as or greater than the likelihoods for all previously monitored days, up to 60 days. We compute a daily empirical P value, p_d, for current day d as follows:

1. Compute $\Pr\left(D(d) \mid \{D(p,d)\}_{p=1}^{pts(d)}, \{\Pr_d(dx_i)\}_{i=0}^{n}\right)$.

2. Compute $\Pr\left(D(d') \mid \{D(p,d')\}_{p=1}^{pts(d')}, \{\Pr_{d'}(dx_i)\}_{i=0}^{n}\right)$ for $d' = max(0, d-60), \ldots, d-1$.

3. Set p_d to the fraction of times the terms computed in step 2 are less than the terms computed in step 1.

That is, we compute the likelihood of the data for the previous (up to) 60 days, then compare the likelihood of the data on day d to those values. We say that day d is unusual if $P_d \leq 0.01$. A sequence of unusual days, each with $P_d \leq 0.01$, may signify an outbreak of a novel disease.

Data and Modeling

Our data come from 5 University of Pittsburgh Medical Center (UPMC) hospitals serving Allegheny County in Southwestern Pennsylvania. As of the 2020 census, the population was approximately 1,223,000. Allegheny County encompasses the City of Pittsburgh which accounts for approximately 25% of the county population, with the remainder of the population being primarily suburban. The racial composition was approximately 75% White, 13% African American, and 12% other (including Native American, Asian, Pacific Islander,

Hispanic, or Latino). The age distribution of the population is approximately 22% younger than the age of 18 years, 9% from 18 to 24 years, 28% from 25 to 44 years, 23% from 45 to 64 years, and 18% who were 65 years of age or older. UPMC serves approximately 60% of the ED visits in Allegheny County.

The training data set consisted of ED encounters at the 5 UPMC hospitals from June 1, 2010, to May 31, 2014, including 815 influenza, 414 RSV, 198 hMPV, 100 PIV that were laboratory-test positive, and 59,428 other visits. In particular, patient encounters with a positive laboratory test for influenza by polymerase chain reaction, direct fluorescent antibody, or viral culture were labeled as influenza. We use similar criteria for labeling patient cases with RSV, hMPV, and PIV. For training purposes, we excluded cases that have positive laboratory results for more than 1 virus. The 59,428 other visits were defined as visits in August 2010, 2011, 2012, or 2013, which did not have any influenza, RSV, hMPV, or PIV laboratory tests performed. These other cases were obtained from visits during August of those 4 years because relatively few outbreaks of these 4 diseases occurred during that month. The testing data set consisted of ED encounters from June 1, 2014, to May 31, 2015.

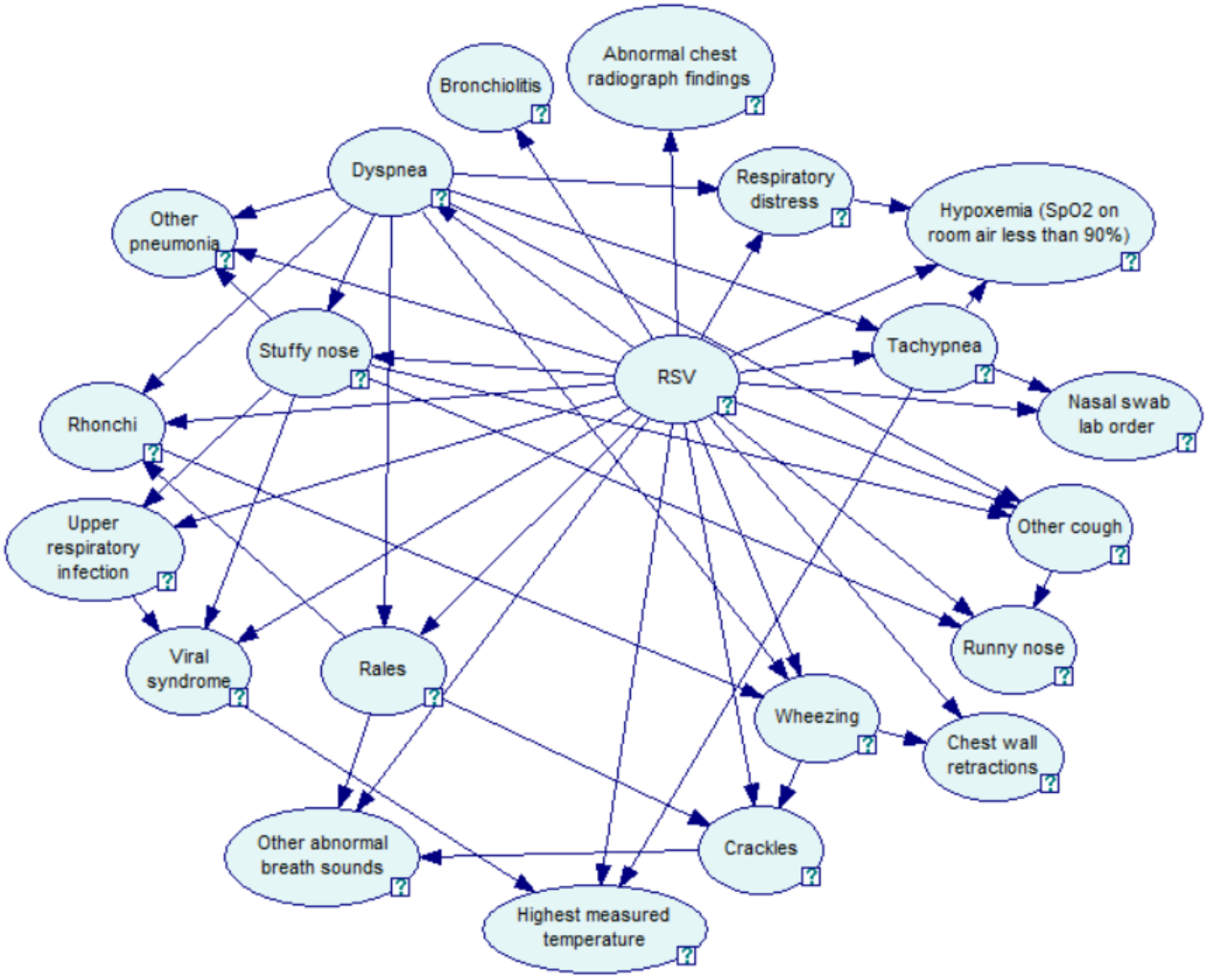
Free-text patient care reports were used as the input to the Topaz parser [32], which for each report generates the status (value) for each of the 79 clinical findings that clinical experts have deemed relevant to ILIs. The highest measured temperature was classified into 4 categories, which included ≥104.0 F (40.0 C), 100.4-103.9 F (38.0-39.9 C), <100.4 F (38.0 C), and unknown. Each of the remaining findings took the values present, absent, or unknown. The designation of “absent” indicated that the clinician had reported the finding as being absent (eg, “patient denies sore throat”). We discarded those findings with information gain scores (regarding disease diagnosis) of zero. Because there is a testing bias for ILIs across age groups, the age finding was not included.

Our approach to outbreak detection is based on modeling and tracking patients with known diseases and noting anomalies. This requires modeling each patient. To that end, we developed 5 Bayesian network disease models (ie, influenza, RSV, hMPV, PIV, and other) using the same search process to find each Bayesian network structure, as follows. We started with a naïve Bayes network structure in which all findings have an arc from the disease node to each finding node. We then used the K2 learning algorithm [33] to identify additional arcs among the clinical feature nodes which were assigned an arbitrary ordering. The search was based on the K2 Bayesian score with a restriction that each finding could have at most 2 parents beyond the disease node. After finding a network structure, we estimated from the data conditional probabilities in the network model. For instance, Figure 1 shows the relationships among the top 20 features of the RSV model [34].

We used the area under the receiver operating characteristic curve (AUC) [35] as the measure of discrimination performance. The AUCs obtained for influenza, RSV, hMPV, PIV, and other were 0.88, 0.92, 0.91, 0.89, and 0.90, respectively. In testing the performance of the influenza model, the disease-positive group consists of patient cases that have positive laboratory results for influenza. The negative test group consists of cases that either have negative laboratory results for influenza or have not had any laboratory tests performed for influenza. It is likely, however, that some of the patient cases without any laboratory tests for influenza will have influenza, which we would expect to have reduced the AUC that we report. An analogous situation exists for testing the RSV, hMPV, and PIV models.

When testing the model of other disease, the negative test group includes cases that have at least 1 positive laboratory test result for influenza, RSV, hMPV, or PIV. The positive test group consists of cases that either have negative laboratory results for all 4 respiratory diseases or have not had any of those tests performed. It is likely, however, that some of the cases without any tests performed will have one or more of the 4 respiratory diseases. Given this consideration and the discussion in the previous paragraph, the reported AUCs are likely to represent lower bounds on performance that would be obtained if the test case labels were more accurate.

Figure 1. A Bayesian network model for the 20 most informative findings in the RSV model. RSV: respiratory syncytial virus; SpO2: measure of the saturation of peripheral blood oxygen.



Ethical Considerations

The research protocol was approved by the University of Pittsburgh Institutional Review Board (study number 20030193). As no patients were enrolled in this study and no compensation was offered, we obtained a waiver of consent from the institutional review board. A UPMC-approved honest broker deidentified the data to meet the requirements of a limited data set before distribution.

Results

This section reports the results we obtained in applying ILI Tracker to the data described above to estimate the presence of the outbreak diseases we modeled over time and to monitor for the presence of a novel disease in the population.

Tracking of Known Diseases

Figure 2 shows the results of running the ILI Tracker for the period June 1, 2014, to May 31, 2015. The red (dashed) lines are the daily number of laboratory-confirmed cases and the blue (solid) lines are the expected number of patients computed by ILI Tracker each day. The ILI-expected case predictions appear to be correlated with the number of positive laboratory results for those diseases. Note, however, that the expected and confirmed cases of RSV before January 1 deviate significantly. The possibility this deviation was due to a novel disease is discussed below. We computed the correlation between the daily number of expected and confirmed cases for each disease. Table 1 shows the Pearson and Spearman *r* and P values. Across the 4 modeled diseases, the peak days predicted by the ILI Tracker were close to the peak days according to the laboratory-confirmed cases. The peak days of the 7-day moving averages differed by 6 days for influenza, 6 days for RSV, 0 days for hMPV, and 4 days for PIV.

Figure 2. Expected and confirmed cases for June 1, 2014, to May 31, 2015 (7-day moving average). hMPV: human metapneumovirus; ILI: influenza-like illness; RSV: respiratory syncytial virus.

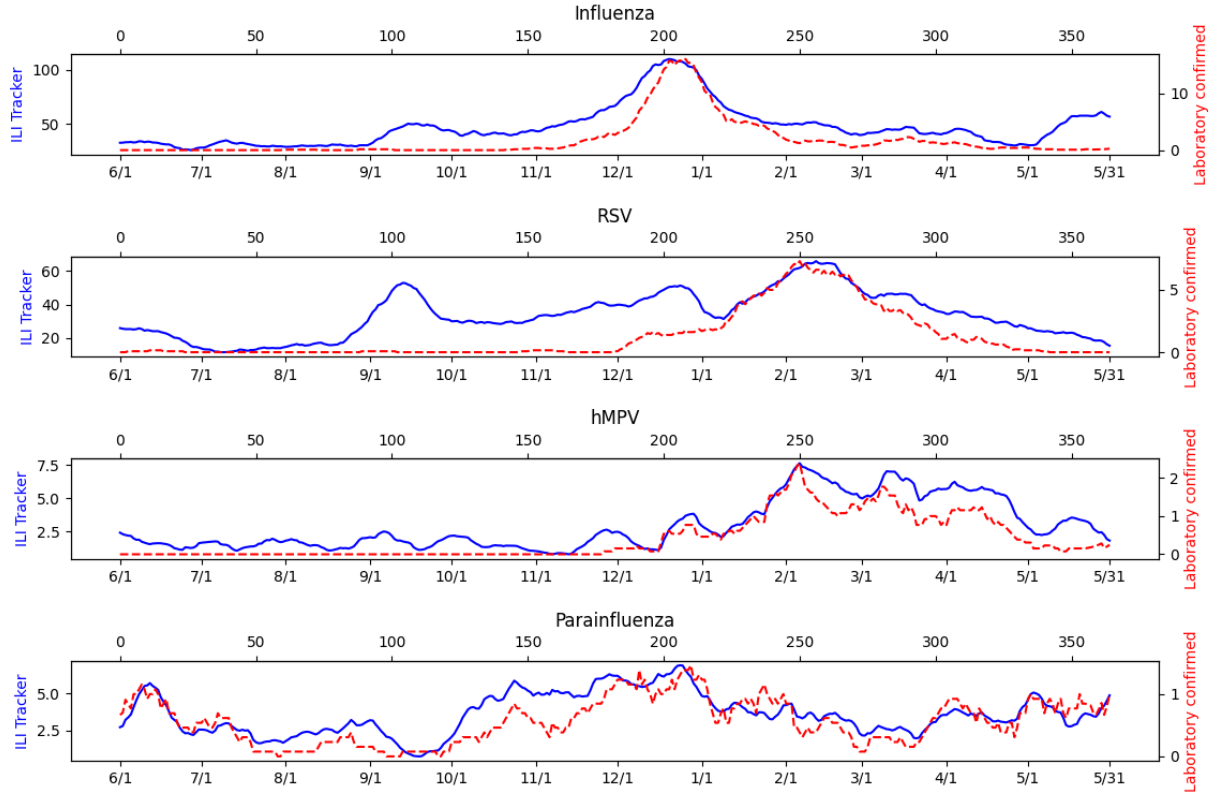


Table 1. Comparison of ILI Tracker and confirmed cases from June 1, 2014, to May 31, 2015, as measured using Pearson and Spearman correlations. The P values are the probability of the r values if the correlations were 0.

Disease	Pearson *r* (P value)	Spearman *r* (P value)
Influenza	0.81 (<.001)	0.63 (<.001)
RSV[a]	0.66 (<.001)	0.64 (<.001)
hMPV[b]	0.72 (<.001)	0.65 (<.001)
PIV[c]	0.51 (<.001)	0.52 (<.001)

[a]RSV: respiratory syncytial virus.

[b]hMPV: human metapneumovirus.

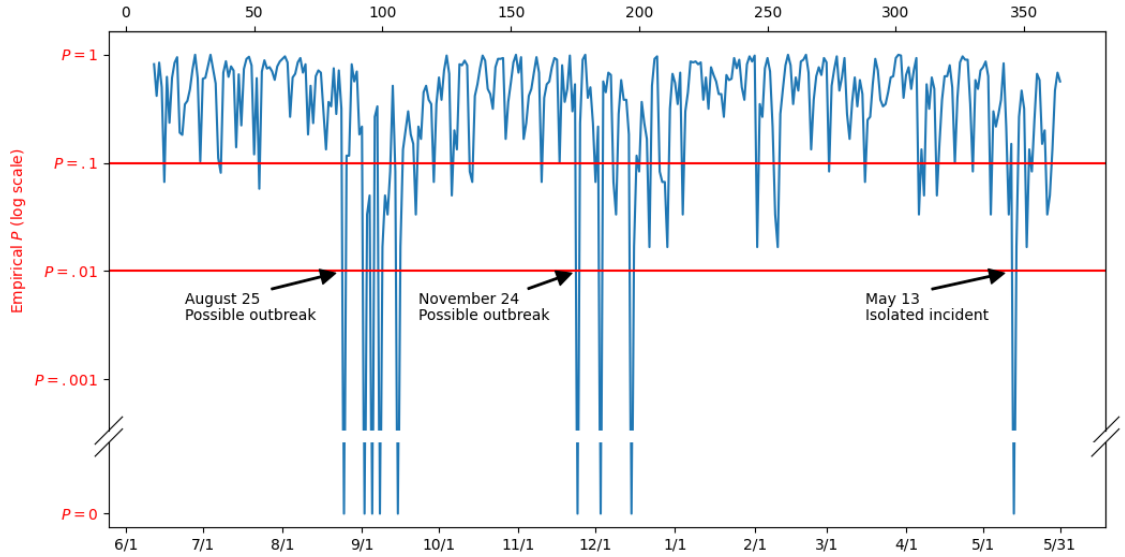
[c]PIV: parainfluenza.

Putative Detection of a Novel Disease

The system reported finding a novel disease which, based on CDC reports for the time period being reported on, appears to be an enterovirus D68 (EV-D68) outbreak.

Figure 3 shows the daily empirical P values from June 1, 2014, to May 31, 2015. The horizontal red lines indicate P=.1 and P=.01. As mentioned above, P=.01 is our threshold for unusual. On August 25, 2014, ILI Tracker signaled an unusual day. Although a single unusual day is not necessarily the beginning of an outbreak, it may warrant further investigation. A total of 8 of the 10 most unlikely patients on August 25 (ie, patients with very low probability findings given the expected prevalence of modeled diseases in the ED) showed signs of a respiratory illness. Starting on September 2, ILI Tracker noted 4 additional unusual days within a single week, with the most unlikely patients again showing signs of a respiratory illness. By September 9, there were sufficient data to characterize these anomalies in terms of the most prevalent findings.

Figure 3. Daily empirical P values from June 1, 2014, to May 31, 2015.



Each day, *d*, we computed the expected number of patients with each finding, *f*, based on the rate of that finding for each disease and the prevalence of each disease in the ED on that day, with the formula $E(f)=\sum_{i=0}^{n} Pr(f \mid dx_i) Pr_d(dx_i)$.

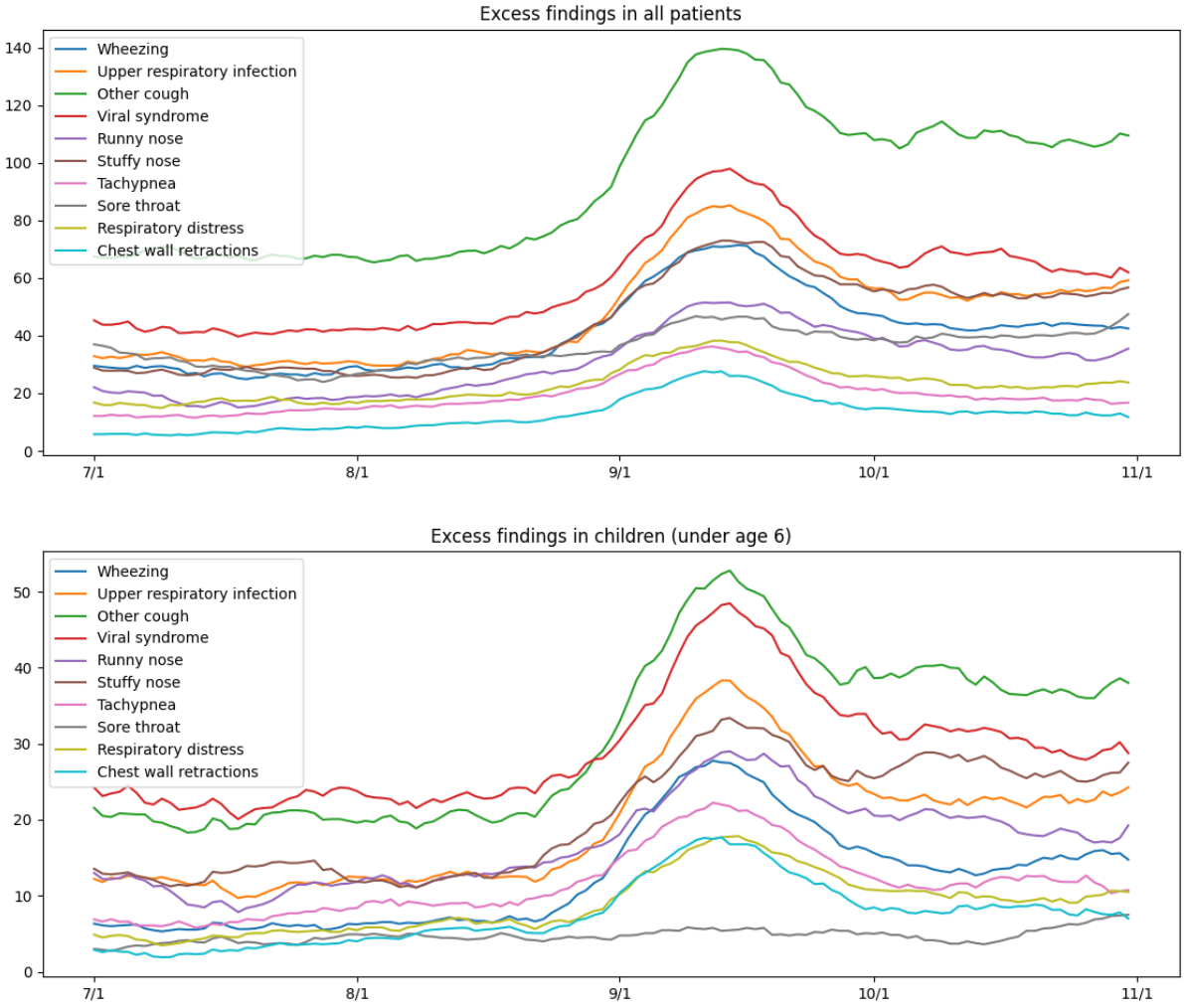
If patients with a novel disease are present in the ED, we expect the actual number of findings in the data that are characteristic of the novel disease to exceed the expected number of those findings when assuming that only modeled diseases are present. We identified the top 10 most excessive findings for one week starting September 2. Figure 4 tracks the daily occurrence of these findings from July to October 2014. There was a clear increase in their frequency starting in the latter part of August.

During the late summer (late August and early September) of 2014, the ED of the UPMC Children's Hospital of Pittsburgh experienced an abrupt increase in children presenting with acute respiratory illness, asthma exacerbation, and dyspnea. While symptoms overlapped with common causes of community-acquired viruses, they were unique based on the severity of illness, the sheer volume of children seeking care, and the timing being unusual for any of the annual common acute respiratory viruses that usually circulate. Rapid testing was negative for influenza and RSV. Even the full nucleic acid–based clinically available assays were confusing as some children tested positive for rhinovirus but were much more ill than expected. The assay was not supposed to cross-react with enterovirus, but it did.

During August and the fall of 2014, the CDC identified an outbreak of EV-D68 in the United States, especially among children [36]. As reported by the New Vaccine Surveillance Network (NVSN) [37], common symptoms of EV-D68 include cough, nasal congestion or rhinorrhea, wheezing, and shortness of breath or dyspnea [38]. These symptoms were among the top 10 excess symptoms identified by the ILI Tracker during the same time period (Figure 4). Although fever is often a common symptom of many ILIs, it was neither a common symptom identified by NVSN in children with EV-D68 nor was it among the excess symptoms identified by ILI Tracker. Neither Topaz nor the NVSN was designed to capture neurologic outcomes, such as acute flaccid myelitis, which is a rare but particularly severe neurologic consequence of EV-D68 leading to paralysis [39]. All these factors support that the novel disease identified by ILI Tracker is EV-D68.

On November 24, 2014, the ILI Tracker again signaled a day with highly unusual patient findings. This was a weaker signal (Figure 3) than the signal in late August and early September. Nine of the 10 most unusual patients showed signs of a respiratory illness. The ILI Tracker also noted an isolated unusual day on May 13, 2015.

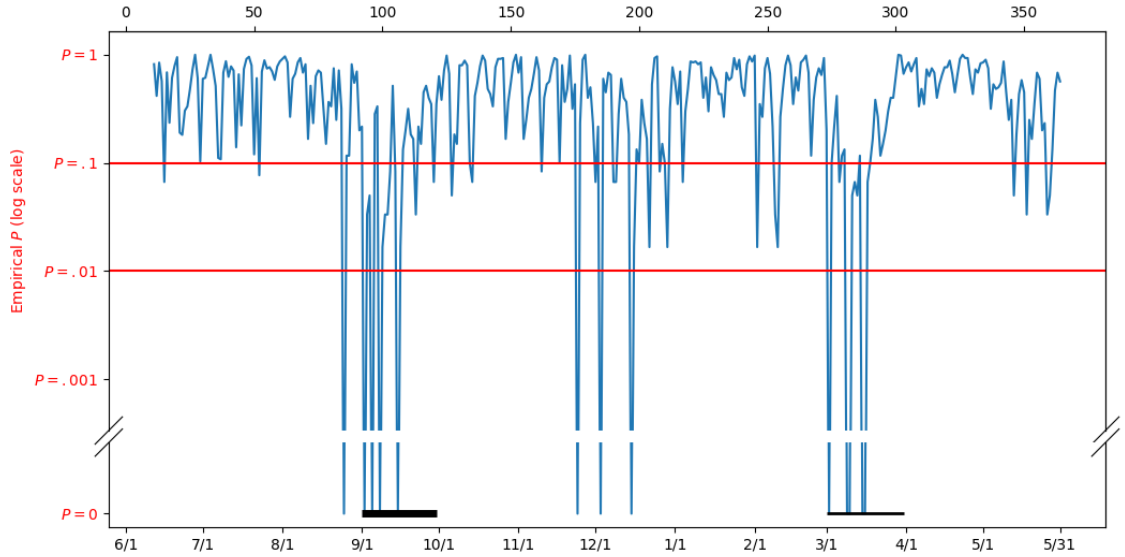
Figure 4. Daily absolute counts of the top 10 excess findings from July to October 2014.



Experiment With a Synthetic Outbreak

ILI Tracker identified a novel outbreak in late August and September of 2014. To further test the ILI Tracker, we created a synthetic outbreak by identifying cases from September 1, 2014, to September 30, 2014, with upper respiratory infection, respiratory distress, or chest wall retractions and artificially adding them from March 1 to 30, 2015. Figure 5 shows the empirical P values computed by ILI Tracker for the outbreak year June 1, 2014, to May 31, 2015, with this new artificial outbreak added to March. The thick black horizontal bars indicate where these cases were copied from (September 1-30) and the thin black horizontal bars indicate where they were copied to (March 1-30). A total of 2787 of these unusual cases from September 2014 were added to March 2015. The results in Figure 5 support that the ILI Tracker was able to identify this "outbreak" despite significant background activity from influenza, RSV, hMPV, and PIV during that time as shown in Figure 2. Note that the ILI Tracker produced a signal by the second day of March at which time only 152 of these cases had entered the ED. Note also that the ILI Tracker did not signal a novel outbreak in March when the cases of the artificial outbreak were not added to the data of that month.

Figure 5. Daily empirical P values from June 1, 2014, to May 31, 2015, with a synthetic outbreak added to March 2015.



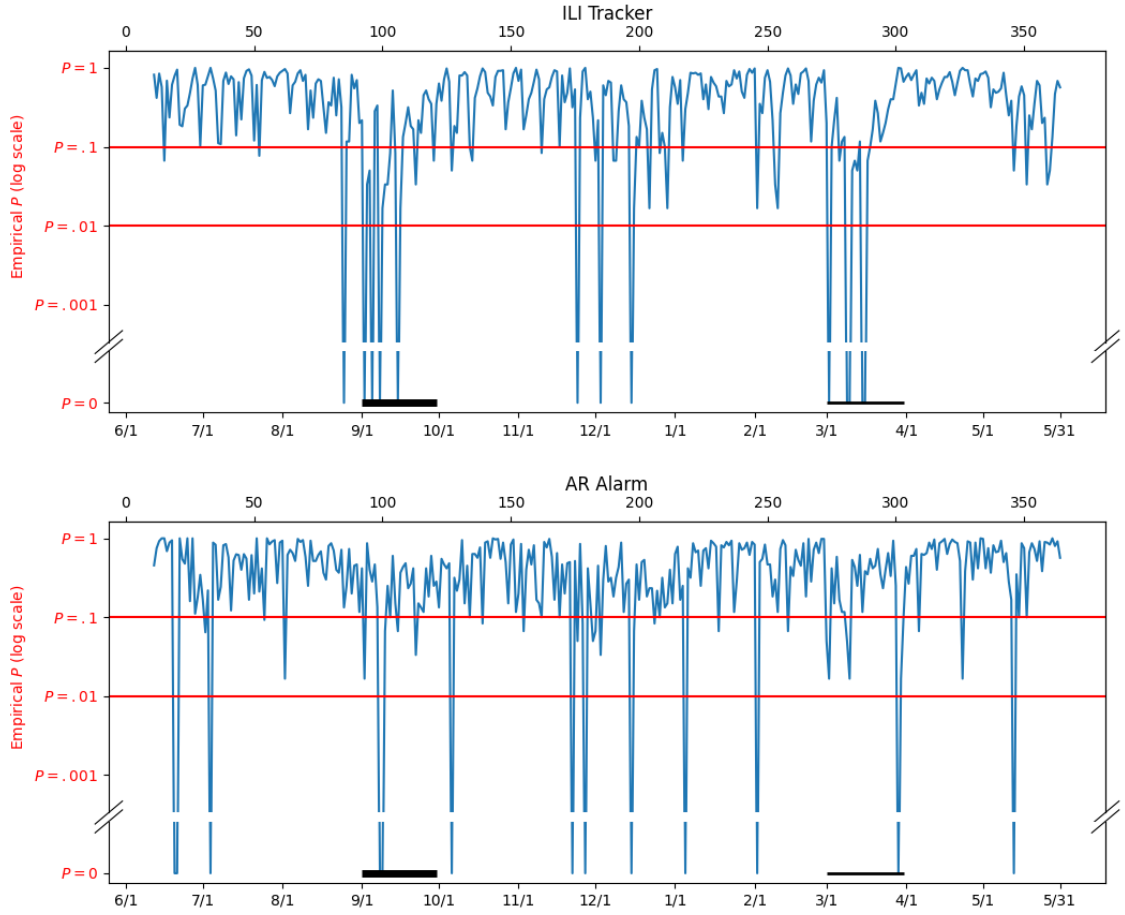
Comparison to an Alternate System

Autoregression is a standard method for outbreak detection [40-42]. To better understand the ability of the ILI Tracker to detect unmodeled outbreaks, we implemented an outbreak system based on autoregression and compared its performance with that of the ILI Tracker. We call the autoregression-based system AR Alarm.

For each finding, *f*, on each day, *d*, we built an AR(7) (autoregression with a lag term for each of the 7 previous days) model using the daily counts for that finding up to (but not including) the current day. We then used the model to predict the number of patients with that finding for the current day and called it *predicted(f,d)*. We then set *p*(*f*,*d*)=*predicted*(*f*,*d*)/*pts*(*d*) where *pts*(*d*) is the total number of patients on day *d*. Thus, *p*(*f*,*d*) is the probability (predicted by the AR(7) model) that an arbitrary patient on day *d* has finding *f*. We then compute the probability (according to the AR(7) model) of seeing the actual number of patients with finding *f* on day *d*, *actual(f,d)*, with the formula *Pr(actual(f,d))=Binom(actual(f,d),pts(d),p(f,d))*, where *Binom* is the binomial distribution. We use the Simple Bayes assumption and set the probability of the set of findings on the current day, *d*, to $\prod_{i=1}^{n} Pr(actual(f_i,d))$, where $\{f_i\}$ is the set of findings. We then compute an empirical P value of the findings on each day using the method described in the "Detecting the Presence of Unmodeled Diseases" section.

Figure 6 compares the empirical P values for ILI Tracker to those from AR Alarm on the data from June 1, 2014, to May 31, 2015, including the artificial outbreak (described above) added to March. After the first 60 days, when the systems are calibrating themselves, they are in general agreement. Note, however, that ILI Tracker notices the August to September outbreak 14 days earlier than AR Alarm and with a stronger signal. Also, AR Alarm produces a few isolated signals, while ILI Tracker produces weaker signals (early October, early January, early February, and late March). Note that the ILI Tracker produces a strong signal and the AR Alarm produces a weaker signal in March when the artificial outbreak was added to the data.

Figure 6. Comparison of empirical P values from ILI Tracker and AR Alarm for June 1, 2014 to May 31, 2015, with a synthetic outbreak added to March 2015. AR: autoregression; ILI: influenza-like illness.



Runtime Analysis

Using a computer with 2 processors, each with six 1.6-GHz cores, it took less than 1 minute to construct all the disease models from the 4 years of training data.

Daily processing occurred in 3 phases. First, feature extraction from ED patient care reports. Second, computation of disease likelihoods for each patient. Third, computation of the expected number of each modeled disease and the P value of the data for each day.

On average, there were about 700 patients each day. Feature extraction typically took about 5 minutes each day using a computer with four 3.6-GHz cores. Computation of disease likelihoods typically took less than 1 minute for the entire set of patients on a given day. Given the likelihood for each patient, the ILI Tracker algorithm takes less than 10 seconds to compute the expected number of each modeled ILI and P value of the data each day using a computer with four 2.5-GHz cores. Thus, the total time to run the ILI Tracker per day on a desktop computer was less than 10 minutes.

We note that feature extraction and computation of disease likelihoods for each patient report are independent of the others. Thus, additional health care facilities can be added to our surveillance system and each can process their patient care reports using their own hardware, which would maintain run-time tractability. The runtime of the ILI Tracker algorithm for a given day is $O(P \times D \times F)$ for each day, where P is the number of patients, D is the number of modeled diseases, and F is the number of modeled findings. Processing time increases linearly as patients, diseases, and features are added, and thus, the algorithm is readily scalable.

Discussion

The performance of ILI Tracker during a 1-year period was moderate for tracking PIV, strong for tracking RSV and hMPV, and very strong for tracking influenza, as measured using Pearson correlation between the tracking and the laboratory-confirmed cases. Using Spearman correlation, ILI Tracker's performance was moderate for tracking PIV and strong for tracking hMPV, RSV, and influenza.

As mentioned above, in late August 2014, ILI Tracker alerted on an outbreak consistent with the EV-D68 outbreak that was identified as present in the United States during that period by the CDC. To our knowledge, that EV-D68 outbreak is the only novel outbreak that was documented to have occurred during the period of our study (June 1, 2014, to May 31, 2015). ILI Tracker is intended to be used as a daily monitor that has its alerts interpreted by clinicians and public health officials. It can identify and output unusual patients for further evaluation by such individuals. Approaches that use aggregate statistics (based on overall counts of findings) cannot identify individual patients who are likely to have an outbreak disease.

If ILI Tracker had been in operation during 2014 it would have signaled a statistical anomaly among patients in late August and provided a set of patients for further investigation. Based on clinical judgment, these patients could be assessed, tested, and possibly isolated. In some cases, samples might be obtained for rapid sequencing. By early September, ILI Tracker could also have provided a preliminary clinical description and timeline of a cohort of unusual patients who turned out to have findings consistent with an outbreak of EV-D68. In the future, such information could help clinicians and public health officials to detect, isolate, characterize, and identify such a novel disease early in the outbreak.

A putative outbreak of an unmodeled disease also appears to have occurred in late November and December of 2014 during outbreaks of both RSV and influenza. Although these patients could easily be lost among a large number of patients with RSV and influenza, they are statistically unlikely to be among those known diseases, according to the analysis by ILI Tracker. Although this was a weak signal, it was statistically significant enough to warrant further investigation. Again, if ILI Tracker had been in operation at that time, it would have identified patients for further evaluation. The source of that putative outbreak remains an open question.

The unusual day detected on May 13, 2015 is likely an isolated incident. Statistically, we should expect such incidents to occur occasionally. ILI Tracker would provide a set of candidate patients for consideration by clinicians and public health officials in the region.

The results reported in this paper provide support that ILI Tracker was able to track well 4 modeled ILI-like diseases over a 1-year period, relative to laboratory-confirmed cases, and it was computationally efficient in doing so. The results we presented also provide support that the system was able to detect a novel outbreak of EV-D68 early in an outbreak that occurred in Allegheny County in 2014, as well as clinically characterize that outbreak disease accurately. Detection was very efficient computationally. In general, the ILI Tracker scales linearly in the number of diseases and number of findings per disease. Thus, it can be expanded to model many additional, known outbreak diseases and their findings.

This work has some important limitations that we plan to address in the future. Here, we assumed a small set of possible (unique) diagnoses. In future work, we plan to extend ILI Tracker to model additional (potentially co-occurring) respiratory diseases, including adenovirus, enterovirus, and SARS-CoV-2. We currently use the Topaz parser to extract a set of less than 100 findings selected specifically for their relevance to ILIs. We plan to use the MetaMap system [43] to expand the set of findings to many thousands that are encoded using Unified Medical Language System concept unique identifiers [44]. In addition, we plan to evaluate ILI Tracker and its extensions on additional years of data from a broader range of hospitals. We implicitly assume that patient care does not vary over time, reporting is constant and comprehensive, and mild cases are noted and documented (regardless of their primary diagnosis).

Our ultimate goal is to deploy an effective, free, and open-source early-warning surveillance system for use in monitoring data in hospital EDs. A preliminary version of the ILI Tracker software will be available on GitHub [45] with updates and test data planned for the future.

Acknowledgments

This work was supported by grant R01LM013509 (Automated Surveillance of Overlapping Outbreaks and New Outbreak Diseases) from the National Library of Medicine of the US National Institutes of Health (NIH). HH was supported by Models of Infectious Disease Agent Study grant U24GM132013. YY was supported by grant R00LM013383 from NIH/National Library of Medicine. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflicts of Interest

JE is President of General Biodefense LLC. All other authors have no conflicts of interest to declare.

References

1. Dato V, Shephard R, Wagner M. Outbreaks and investigations. In: Wagner MM, Moore AW, Aryel RM, editors. Handbook of Biosurveillance. Cambridge, MA. Elsevier Academic Press; 2006:13-26.
2. Wagner MM, Gresham LS, Dato V. Case detection, outbreak detection, and outbreak characterization. In: Wagner MM, Moore AW, Aryel RM, editors. Handbook of Biosurveillance. Cambridge, MA. Academic Press; 2006:27-50.
3. Velikina R, Dato V, Wagner MM. Governmental public health. In: Wagner MM, Moore AW, Aryel RM, editors. Handbook of Biosurveillance. Cambridge, MA. Academic Press; 2006:67-88.

4. Wagner MM, Hogan WR, Aryel RM. The healthcare system. In: Wagner MM, Moore AW, Aryel RM, editors. Handbook of Biosurveillance. Cambridge, MA. Academic Press; 2006:89-110.
5. Brokopp C, Resultan E, Holmes H, Wagner MM. Laboratories. In: Wagner MM, Moore AW, Aryel RM, editors. Handbook of Biosurveillance. Cambridge, MA. Academic Press; 2006:129-142.
6. Metcalf CJE, Lessler J. Opportunities and challenges in modeling emerging infectious diseases. Science. 2017;357(6347):149-152. [FREE Full text] [doi: 10.1126/science.aam8335] [Medline: 28706037]
7. Holmes EC, Rambaut A, Andersen KG. Pandemics: spend on surveillance, not prediction. Nature. 2018;558(7709):180-182. [doi: 10.1038/d41586-018-05373-w] [Medline: 29880819]
8. Villanueva J, Schweitzer B, Odle M, Aden T. Detecting emerging infectious diseases: an overview of the laboratory response network for biological threats. Public Health Rep. 2019;134(2_suppl):16S-21S. [FREE Full text] [doi: 10.1177/0033354919874354] [Medline: 31682559]
9. Smith G, Hippisley-Cox J, Harcourt S, Heaps M, Painter M, Porter A, et al. Developing a national primary care-based early warning system for health protection--a surveillance tool for the future? Analysis of routinely collected data. J Public Health (Oxf). 2007;29(1):75-82. [doi: 10.1093/pubmed/fdl078] [Medline: 17158478]
10. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
11. 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. PLoS Comput Biol. 2013;9(10):e1003256. [FREE Full text] [doi: 10.1371/journal.pcbi.1003256] [Medline: 24146603]
12. Butler D. When google got flu wrong. Nature. 2013;494(7436):155-156. [doi: 10.1038/494155a] [Medline: 23407515]
13. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of google flu: traps in big data analysis. Science. 2014;343(6176):1203-1205. [doi: 10.1126/science.1248506] [Medline: 24626916]
14. Kandula S, Shaman J. Reappraising the utility of google flu trends. PLoS Comput Biol. 2019;15(8):e1007258. [FREE Full text] [doi: 10.1371/journal.pcbi.1007258] [Medline: 31374088]
15. Chan EH, Sahai V, Conrad C, Brownstein JS. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS Negl Trop Dis. 2011;5(5):e1206. [FREE Full text] [doi: 10.1371/journal.pntd.0001206] [Medline: 21647308]
16. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using google search data via ARGO. Proc Natl Acad Sci USA. 2015;112(47):14473-14478. [FREE Full text] [doi: 10.1073/pnas.1515373112] [Medline: 26553980]
17. van de Belt TH, van Stockum PT, Engelen LJLPG, Lancee J, Schrijver R, Rodríguez-Baño J, et al. Social media posts and online search behaviour as early-warning system for MRSA outbreaks. Antimicrob Resist Infect Control. 2018;7:69. [FREE Full text] [doi: 10.1186/s13756-018-0359-4] [Medline: 29876100]
18. Dai Y, Wang J. Identifying the outbreak signal of COVID-19 before the response of the traditional disease monitoring system. PLoS Negl Trop Dis. 2020;14(10):e0008758. [FREE Full text] [doi: 10.1371/journal.pntd.0008758] [Medline: 33001985]
19. Villamarín R, Cooper G, Wagner M, Tsui FC, Espino JU. A method for estimating from thermometer sales the incidence of diseases that are symptomatically similar to influenza. J Biomed Inform. 2013;46(3):444-457. [FREE Full text] [doi: 10.1016/j.jbi.2013.02.003] [Medline: 23501015]
20. Thacker SB, Choi K, Brachman PS. The surveillance of infectious diseases. JAMA. 1983;249(9):1181-1185. [Medline: 6823080]
21. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health. 2020;41:101-118. [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
22. Donaldson AL, Hardstaff JL, Harris JP, Vivancos R, O'Brien SJ. School-based surveillance of acute infectious disease in children: a systematic review. BMC Infect Dis. 2021;21(1):744. [FREE Full text] [doi: 10.1186/s12879-021-06444-6] [Medline: 34344304]
23. Muleym D, Shanin M, Dias C, Abdullah M. Role of transport during outbreak of infectious diseases: evidence from the past. Sustainability, MDPI. 2020;12(18):7367. [FREE Full text] [doi: 10.3390/su12187367]
24. Simonsen L, Gog JR, Olson D, Viboud C. Infectious disease surveillance in the big data era: towards faster and locally relevant systems. J Infect Dis. 2016;214(suppl_4):S380-S385. [FREE Full text] [doi: 10.1093/infdis/jiw376] [Medline: 28830112]
25. Hughes HE, Edeghere O, O'Brien SJ, Vivancos R, Elliot AJ. Emergency department syndromic surveillance systems: a systematic review. BMC Public Health. 2020;20(1):1891. [FREE Full text] [doi: 10.1186/s12889-020-09949-y] [Medline: 33298000]
26. National Syndromic Surveillance Program. Centers for Disease Control and Prevention. URL: http://www.cdc.gov/nssp/ [accessed 2024-06-07]
27. Li M, Loschen W, Deyneka L, Burkom H, Ising A, Waller A. Time of arrival analysis in NC DETECT to find clusters of interest from unclassified patient visit records. Online J Public Health Inform. 2013;5(1):e61176. [FREE Full text] [doi: 10.5210/ojphi.v5i1.4512]

28. Burkom H, Elbert Y, Piatko C, Fink C. A term-based approach to asyndromic determination of significant case clusters. Online J Public Health Inform. 2015;7(1):e61504. [FREE Full text] [doi: 10.5210/ojphi.v7i1.5675]
29. Nobles M, Lall R, Mathes RW, Neill DB. Presyndromic surveillance for improved detection of emerging public health threats. Sci Adv. 2022;8(44):eabm4920. [FREE Full text] [doi: 10.1126/sciadv.abm4920] [Medline: 36332014]
30. Aronis JM, Ferraro JP, Gesteland PH, Tsui F, Ye Y, Wagner MM, et al. A Bayesian approach for detecting a disease that is not being modeled. PLoS One. 2020;15(2):e0229658. [FREE Full text] [doi: 10.1371/journal.pone.0229658] [Medline: 32109254]
31. Tsui F, Wagner M, Cooper G, Que J, Harkema H, Dowling J, et al. Probabilistic case detection for disease surveillance using data in electronic medical records. Online J Public Health Inform. 2011;3(3):ojphi.v3i3.3793. [FREE Full text] [doi: 10.5210/ojphi.v3i3.3793] [Medline: 23569615]
32. Chapman WW, Harkema H. Identifying respiratory-related clinical conditions from ED reports with Topaz. Clinical Medicine & Research. 2010;8(1):53. [doi: 10.3121/cmr.8.1.53-b]
33. Cooper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn. 1992;9(4):309-347. [FREE Full text] [doi: 10.1007/bf00994110]
34. Druzdzel MJ. GeNIe: a development environment for graphical decision-analytic models. Philadelphia, PA. Hanley & Belfus; 1999. Presented at: Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association; November 6-10:1206; Washington, DC. URL: https://sites.pitt.edu/~druzdzel/psfiles/amia99.pdf
35. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
36. Midgley CM, Jackson MA, Selvarangan R, Turabelidze G, Obringer E, Johnson D, et al. Severe respiratory illness associated with enterovirus D68 - Missouri and Illinois, 2014. MMWR Morb Mortal Wkly Rep. 2014;63(36):798-799. [FREE Full text] [Medline: 25211545]
37. Perez A, Lively JY, Curns A, Weinberg GA, Halasa NB, Staat MA, et al. New Vaccine Surveillance Network Collaborators. Respiratory virus surveillance among children with acute respiratory illnesses - new vaccine surveillance network, United States, 2016-2021. MMWR Morb Mortal Wkly Rep. 2022;71(40):1253-1259. [FREE Full text] [doi: 10.15585/mmwr.mm7140a1] [Medline: 36201373]
38. Shah MM, Perez A, Lively JY, Avadhanula V, Boom JA, Chappell J, et al. Enterovirus d68-associated acute respiratory illness - new vaccine surveillance network, United States, July-November 2018-2020. MMWR Morb Mortal Wkly Rep. 2021;70(47):1623-1628. [FREE Full text] [doi: 10.15585/mmwr.mm7047a1] [Medline: 34818320]
39. Aliabadi N, Messacar K, Pastula DM, Robinson CC, Leshem E, Sejvar JJ, et al. Enterovirus D68 infection in children with acute flaccid myelitis, Colorado, USA, 2014. Emerg Infect Dis. Aug 2016;22(8):1387-1394. [FREE Full text] [doi: 10.3201/eid2208.151949] [Medline: 27434186]
40. Briët OJT, Amerasinghe PH, Vounatsou P. Generalized seasonal autoregressive integrated moving average models for count data with application to malaria time series with low case numbers. PLoS One. 2013;8(6):e65761. [FREE Full text] [doi: 10.1371/journal.pone.0065761] [Medline: 23785448]
41. Zhang X, Zhang T, Young AA, Li X. Applications and comparisons of four time series models in epidemiological surveillance data. PLoS One. 2014;9(2):e88075. [FREE Full text] [doi: 10.1371/journal.pone.0088075] [Medline: 24505382]
42. Rao Y, McCabe B. Real-time surveillance for abnormal events: the case of influenza outbreaks. Stat Med. 2016;35(13):2206-2220. [doi: 10.1002/sim.6857] [Medline: 26782751]
43. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. 2001. Presented at: American Medical Informatics Association Annual Symposium; November 3-7; Washington, DC. URL: https://europepmc.org/abstract/MED/11825149
44. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267-D270. [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
45. RODS Laboratory GitHub. URL: https://github.com/RodsLaboratory [accessed 2024-06-07]

Abbreviations

AR(7): autoregression with a lag term for each of the 7 previous days
AUC: area under the receiver operating characteristic curve
CDC: Centers for Disease Control and Prevention
ED: emergency department
EV-D68: enterovirus D68
hMPV: human metapneumovirus
ILI: influenza-like illness
NVSN: New Vaccine Surveillance Network
PIV: parainfluenza
RSV: respiratory syncytial virus
UPMC: University of Pittsburgh Medical Center

Edited by A Mavragani; submitted 14.02.24; peer-reviewed by G Weber, A Couture; comments to author 09.04.24; revised version received 02.05.24; accepted 24.05.24; published 13.08.24

Please cite as:
Aronis JM, Ye Y, Espino J, Hochheiser H, Michaels MG, Cooper GF
A Bayesian System to Detect and Track Outbreaks of Influenza-Like Illnesses Including Novel Diseases: Algorithm Development and Validation
JMIR Public Health Surveill 2024;10:e57349
URL: https://publichealth.jmir.org/2024/1/e57349
doi: 10.2196/57349
PMID: 38805611