Journal of Biomedical Informatics 53 (2015) 180–187

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Quantifying the determinants of outbreak detection performance through simulation and machine learning

Nastaran Jafarpour [a,*], Masoumeh Izadi [b], Doina Precup [c], David L. Buckeridge [b]

[a] Department of Computer Engineering, Ecole Polytechnique de Montreal, C.P. 6079, succursale Centre-ville, Montreal, Quebec H3C 3A7, Canada
[b] Department of Epidemiology and Biostatistics, McGill University, Clinical and Health Informatics Research Group, 1140 Pine Ave. West, Montreal, Quebec H3A 1A3, Canada
[c] School of Computer Science, McGill University, 3480 University St., Montreal, Quebec H3A 0E7, Canada

ARTICLE INFO

Article history:
Received 4 July 2014
Accepted 27 October 2014
Available online 6 November 2014

Keywords:
Disease outbreak detection
Surveillance
Bayesian networks
Predicting performance
Public health informatics
Outbreak simulation

ABSTRACT

Objective: To develop a probabilistic model for discovering and quantifying determinants of outbreak detection and to use the model to predict detection performance for new outbreaks.
Materials and methods: We used an existing software platform to simulate waterborne disease outbreaks of varying duration and magnitude. The simulated data were overlaid on real data from visits to emergency department in Montreal for gastroenteritis. We analyzed the combined data using biosurveillance algorithms, varying their parameters over a wide range. We then applied structure and parameter learning algorithms to the resulting data set to build a Bayesian network model for predicting detection performance as a function of outbreak characteristics and surveillance system parameters. We evaluated the predictions of this model through 5-fold cross-validation.
Results: The model predicted performance metrics of commonly used outbreak detection methods with an accuracy greater than 0.80. The model also quantified the influence of different outbreak characteristics and parameters of biosurveillance algorithms on detection performance in practically relevant surveillance scenarios. In addition to identifying characteristics expected *a priori* to have a strong influence on detection performance, such as the alerting threshold and the peak size of the outbreak, the model suggested an important role for other algorithm features, such as adjustment for weekly patterns.
Conclusion: We developed a model that accurately predicts how characteristics of disease outbreaks and detection methods will influence on detection. This model can be used to compare the performance of detection methods under different surveillance scenarios, to gain insight into which characteristics of outbreaks and biosurveillance algorithms drive detection performance, and to guide the configuration of surveillance systems.

1. Introduction

The past decade has seen the emergence of diseases caused by previously unrecognized threats and the sudden appearance of known diseases in new environments. Consequently, infectious diseases continue to cause high human and financial costs. In order to prevent the spread of infectious diseases, early detection of disease outbreaks is crucial. One approach to early detection is to use automated syndromic surveillance systems, which monitor health-related data from different sources to detect potential disease outbreaks in a timely fashion.

Syndromic surveillance systems continuously apply statistical algorithms to large volumes of data generated through health-related behaviors (e.g. counts of emergency department visits) to detect anomalies and support investigation and control measures.

Many outbreak detection algorithms have been proposed for use in syndromic surveillance. While it is clear that algorithms perform differently when they are applied to different data sources or used in different surveillance situations, there is insufficient empirical evidence regarding the effectiveness of algorithms under different conditions. Such evidence could guide public health practitioners in the choice of surveillance systems algorithms and configurations. The few existing studies evaluating detection performance are based on data that are not publicly available, making evaluations difficult to generalize or replicate [1]. Moreover, the performance of detection algorithms is influenced by many factors, including the nature of the disease, characteristics of the outbreak

* Corresponding author.
E-mail addresses: nastaran.jafarpour@polymtl.ca (N. Jafarpour), mtabae@cs.mcgill.ca (M. Izadi), dprecup@cs.mcgill.ca (D. Precup), david.buckeridge@mcgill.ca (D.L. Buckeridge).

http://dx.doi.org/10.1016/j.jbi.2014.10.009
1532-0464/

signal (such as peak size and intensity), baseline data (such as weekly mean and standard deviation), and parameters of the detection method used (such as alerting threshold). Some researchers [2] argue that the lack of a standardized framework for the assessment of outbreak detection methods and the diversity of factors that influence detection performance decreases the ability to compare detection methods.

The objective of this research is to develop and evaluate a model for quantitatively characterizing the determinants of outbreak detection performance and predicting the performance of detection methods. Earlier work [3] showed that it is possible to predict outbreak detection performance quantitatively with acceptable accuracy. That research developed a prediction model based on logistic regression, which assumes a multiplicative relationship between variables. While this model predicted the detection performance of the algorithms with reasonable accuracy, it could not model complex dependencies between variables and their relationship with multiple performance metrics. This limitation was due mainly to the nature of logistic regression, which implements a flat, linear model. In previous work [4], we assessed the feasibility of addressing this limitation by developing a Bayesian network model using data generated thorough simulation. Many different algorithms could be used to model detection performance, such as support vector machines (SVM) and random forests. However, we chose to use a graphical model because it has the advantage of not only providing a prediction of performance, but also providing a representation of the different probabilistic dependencies between outbreak and algorithm characteristics, on one hand, and performance, on the other hand. This information can be useful when trying to understand which factors influence the ability of an outbreak detection algorithm to detect a type of outbreak accurately and in a timely manner.

This paper significantly advances our prior work by combining outbreak data generated by a realistic simulation model with real healthcare utilization data and then evaluating the performance of a wider range of commonly used biosurveillance algorithms. The resulting dataset is used to build and evaluate a Bayesian network model for predicting detection performance. The developed Bayesian network can be used for predicting how well different outbreak detection methods will perform under different circumstances. We illustrate a variety of outbreak scenarios and use inference in the learned Bayesian network to find the best settings for detection methods and predict the detection performance in those scenarios.

While the Bayesian network is built using simulation data, it has two major advantages as a predictor over simply querying the simulation results. On one hand, the Bayesian network is efficient to query when new algorithms or scenarios need to be tested (as opposed to running an expensive simulation). On the other hand, the Bayesian network generalizes the information from the simulation data, allowing queries for outbreak characteristics and surveillance algorithm traits that have not been simulated. A secondary effect of the generalization is to smooth out noise and possible outliers in the simulation data.

The proposed framework for performance evaluation of outbreak detection methods under a wide variety of outbreak circumstances is general and can be used in further studies. We note that while we use the SnAP simulation platform developed at McGill, the same methodology can be used with other count data, provided through alternative simulation methods. We anticipate that the model for predicting detection performance can be used to develop new biosurveillance methods by identifying ideal algorithms characteristics, which may not exist in any currently available algorithms. However, building a new detection method is beyond the scope of this paper.

The structure of this paper is as follows: in Section 2, we review the outbreak detection methods used in our study and describe common measures of detection performance. In Section 3, we describe our simulated surveillance data for waterborne disease outbreaks used in this study and the development and evaluation of our Bayesian network model. In Section 4, we present the accuracy of our model for predicting detection performance and we illustrate its capability to identify factors that influence outbreak detection performance. We also present examples of how the model can be used in practical scenarios. We close with a discussion of the results, concluding remarks, and directions for future work.

2. Background

In public health practice, many approaches are used analyze time series of healthcare utilization records with the goal of detecting disease outbreaks. In this paper, we use a popular set of detection methods based on statistical process control charts, the C-family of detection algorithms [5] and Adaptive Poisson Regression. C1, C2, and C3 are adaptive algorithms included in the Early Aberration Reporting System (EARS) developed by the Centre of Disease Control and Prevention (CDC). The C-algorithms assume that the expected value of the time series for the given time t is the mean of the values observed in a sliding window. If the difference between the observed value at a given time t and the mean of the window divided by the standard deviation of the window is bigger than a *threshold*, an unusual event is flagged and the possibility of a disease outbreak is signaled.

The C-algorithms are distinguished by the configuration of two parameters: the *guardband* and the *memory*. Gradually increasing outbreaks can bias the test statistic upward, so the detection algorithm may fail to flag the outbreak. To avoid this situation, C2 and C3 use a 2-day gap, called a guardband, between the sliding window and the test interval. C3 includes 2 recent observations in the computation of test statistic at time t, which is called *memory*. In the EARS system, the size of the window used for the calculation of the expected value is 7 days; however, this parameter can be varied. Detection algorithms can be configured using various alerting thresholds, which result in different sensitivity and false alarm rates.

Most surveillance tasks based on health care utilization are affected by weekly patterns. Many health-care facilities have fewer visits during weekends and there is a sharp increase in the number of visits on Mondays, which should not be considered an outbreak. The W2 algorithm is a modified version of C2 that takes weekly patterns into account [6], by stratifying the baseline data into two distinct baselines: one for weekdays, the other for weekends. The W3 algorithm is the similar counterpart of the C3 algorithm.

Another outbreak detection method, called Adaptive Poisson Regression, assumes that the distribution of health care utilization counts in the surveillance time series is Poisson and uses categorical variables to represent trends and patterns. Xing described Adaptive Poisson Regression, which uses a sliding window of 56 days for estimating the regression coefficients and alerting *threshold* [7]. The logarithmic link function estimates the expected value at time t as:

$$\log(Expected_t) = c_0 + [c_1 \times dow_{baseline}(t)] + [c_2 \times 14day_{baseline}(t)]$$

where c_0 is a constant intercept, the term $[c_1 \times dow_{baseline}(t)]$captures the day-of-week effect, and the term $[c_2 \times 14day_{baseline}(t)]$ represents the current seasonal trends in cycles of 14 days. The Poisson regression algorithm is adaptive to recent changes in the data and the algorithm parameters. A 2-day *guardband* can be used to avoid the contamination of the sliding window and test interval.

We note that many other surveillance methods can be employed, such as the Shewhart control Chart [8], CUSUM [9], Exponential Weighted Moving Average algorithms (EWMA) [10], the Shiryaew-Roberts method [11], and likelihood ratio-based methods [12]. More details on these methods can be found in [13,14]. Our work does not include these algorithms at the moment, because doing an exhaustive study would be too time consuming. As a result, we picked approaches that seem to be used most extensively in practical surveillance settings. However, it would be conceptually straightforward to extend our framework to include other biosurveillance algorithms and their characteristics.

The performance of outbreak detection algorithms is evaluated in terms of the *specificity* and *timeliness* of detection. Specificity is the probability that no alert will be given when no outbreak has occurred [15]. It is calculated as:

$$specificity = P(alarm = 0 | outbreak = 0) = \frac{n(alarm = 0,\ outbreak = 0)}{n(outbreak = 0)}$$

where $n(alarm = 0, outbreak = 0)$ is the number of non-outbreak days in which the algorithm does not raise an alarm and $n(outbreak = 0)$ is the number of non-outbreak days in an analysis interval.

Timeliness is the proportion of time saved by detection relative to the onset of an outbreak (t_{onset}). If an outbreak is detected, timeliness is defined as:

$$timeliness = 1 - \frac{t_{detection} - t_{onset}}{outbreakDuration}$$

where *outbreakDuration* is the number of days for which outbreak cases occur. The $t_{detection}$ is the index of the day within the time series when the outbreak is detected and t_{onset} is the index of the day on which the outbreak starts. The proportion of delay is subtracted from 1, so higher values denote an earlier detection of the outbreak. Timeliness is 1 if the outbreak is detected on the first day and 0 when the outbreak is not detected at all [15].

In this paper, we measure the *detection* rate as a performance metric. *Detection* is a binary variable which indicates whether each outbreak is detected or not. It can be defined as *sensitivity per outbreak* where sensitivity is the probability of raising an alarm given that an outbreak occurred:

$$sensitivity = P(alarm = 1 | outbreak = 1) = \frac{n(alarm = 1,\ outbreak = 1)}{n(outbreak = 1)}$$

As there is only one outbreak in every time series generated by the simulator, the *detection* will be 1 if the outbreak is detected and 0 if the algorithm does not trigger any alert. We focus on *detection* rate because it measures the overall ability of algorithms to detect different types of outbreaks. We are interested in predicting how the detection rate changes in different surveillance settings and under different algorithm parameter values.

3. Methods

Using administrative data and simulated outbreaks, we created a data set for performance benchmarking of a number of detection algorithms and used a Bayesian network to model the performance of these algorithms.

3.1. Simulated surveillance data

We used a validated model for simulating waterborne outbreaks of cryptosporidiosis [16], the Simulation Analysis Platform (SnAP) [17], to generate surveillance data for this study. The simulation model includes components to represent water distribution, human mobility, exposure to drinking water, infection, disease progression, healthcare utilization, laboratory testing, and reporting to public health. We performed many simulations of surveillance data that would result from a waterborne outbreak due to the failure of a water treatment plant in an urban area. This model creates a synthetic population from census data, and then uses 30 parameters to define the progression of individuals through the model. In the simulation scenarios for generating our data, two parameters were varied systematically: the duration of water contamination, which was varied over 6 values (72, 120, 168, 240, 360 and 480 h), and the cryptosporidium concentration, which was varied over 3 levels (10^{-6}, 10^{-5}, 10^{-4}). The possible combinations of these values define 18 different scenarios. Each of these 18 scenarios was run 1000 times using Latin Hypercube Sampling to randomly select values from hyper-distributions for the other parameters in the model [18]. The simulation parameters and additional details about the SnAP platform are available from the authors of the paper upon request.

The outbreak signals were superimposed on baseline data, which were daily counts of people visiting emergency departments in Montreal for gastroenteritis, over 6 years. The onset of the outbreak was selected randomly, relative to the baseline. Each simulation contains exactly one outbreak. We did not have "ground truth" for the real health care utilization data, which means that real outbreaks could have occurred during this interval. However, there were no known such outbreaks.

3.2. Algorithm benchmarking data

In this paper, we considered the following set of widely used detection algorithms: C1, C2, C3, W2, W3 and Adaptive Poisson Regression with and without guardband. We generated a data set of detection results by applying these algorithms with different parameter values to the surveillance data generated according to the protocol described above. Table 1 presents the features of this data set grouped according to: the parameters of the detection algorithm (memory, guardband, weekly pattern, threshold, and history), the characteristics of the GI baseline data (mean and standard deviation of the number of emergency department (ED) visits over the most recent seven days), the characteristics of the outbreaks added to the baseline data (peak size, time to peak, interval of outbreak days, contamination level and duration of contamination), and the metrics used to measure the performance of a detection algorithm (detection, specificity, and timeliness). This data set contained 72,000 instances, consisting of evaluating each algorithm-parameter combination on each of the 18,000 time series. For each detection algorithm we measured the specificity and timeliness of detection on each of the time series. We also measured *detection* as a binary variable indicating, for each time series, whether the outbreak was detected or not.

The binary variable "weekly pattern" was used as a proxy to indicate whether an algorithm adjusted for day-of-week variations in counts, with zero indicating C1, C2 and C3, and 1 indicating W2, W3, and adaptive Poisson. The variable *sliding window* indicates the size of the window used to calculate the expected value of the ED visit count. Its value is 7 days for the C and W algorithms and 56 days for Adaptive Poisson Regression. The baseline characteristics are statistical characteristics of the ED visit time series without outbreaks. In the data pre-processing step, continuous variables (e.g. *peak size*) were discretized using the k-means function in Netica software [19] for the ease of use in our Bayesian network model.

3.3. Bayesian network model of algorithm characterization

The extensive simulation and algorithm evaluation that we performed requires resources and expertise not available in many

Table 1
Features of algorithm benchmarking data.

Data feature type	Source of data	Data feature	Description	Value
Algorithm parameters	Experimentally defined	Memory	Number of days over which the test statistic is pooled	0, 2
		Guardband	Gap days between the sliding window and the test day	0, 2
		Weekly pattern	Whether or not the algorithm adjusts for a weekly pattern	0, 1
		Threshold	Alerting threshold	0, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5
		Sliding window (history)	Size of the window used for the calculation of the expected value	7, 56
Baseline characteristics	Healthcare utilization	Mean 7	Average data counts in last 7 days	[400, 932]
		SD7	Standard deviation of 7 recent days	[159, 417]
Outbreak characteristics	Simulation	Peak	Peak size as the number of additional counts of outbreak signal above the baseline	[3, 7845]
		Time to peak	Number of days from the onset of the signal to the peak day	[2, 26]
		Outbreak interval	Length of outbreak signal	[4, 52]
		Contamination level	Cryptosporidium concentration in water	10^{-6}, 10^{-5}, 10^{-4}
		Contamination duration	Duration of days of water contamination	72, 120, 168, 240, 360, 480
Detection performance metrics	Performance evaluation of algorithms	Detection	Whether or not the outbreak is detected	0, 1
		Specificity	Probability of no alert when there is no outbreak	[0, 1]
		Timeliness	Proportion of saved time to the outbreak duration	[0, 1]

public health settings. To make the insights from our analysis more generally accessible, and to allow identification of promising new biosurveillance algorithms, we developed a probabilistic model of the detection performance of biosurveillance algorithms for different scenarios. We used Bayesian networks (BNs) to model the relationships between detection performance, algorithm parameters, and outbreak characteristics. BNs capture conditional independence relationships between different variables through a parameterized directed graph [20], and provide a tool for making inference in the form of what-if analysis. As such, they provide a rich formalism for analyzing complex multi-variate data, in which one is interested in uncovering relationships between different variables, rather than just predicting a given type of outcome. A Bayesian network is a directed acyclic graph, where each node represents a random variable or a group of random variables, and the edges express conditional dependencies between random variables[1]. Each node has a conditional probability distribution that reflects the probability of any values for that node given the values of its parents (i.e., the nodes with direct arcs into it). Conditional probability distributions are often represented by tables or trees, which are considered the parameters of a BN model. The relationships in a Bayesian Network model need not to be causal i.e., a directed edge between two nodes does not mean that they are causally dependent [21]. The edges in the network represent the conditional dependencies observed among variables. The network graph structure and its parameters are learned from data using an optimization-based search method that tries to maximize the likelihood function over possible network configurations. We experimented constructing a Bayesian network using several structure learning methods, including Navie Bayes [22], Tree Augmented Naive Bayes, Maximum Spanning Tree, Markov Blanket learning, and Taboo search [23]. We selected the structure with the best prediction performance. After selecting the structure, the Netica software package version 5.08 [19] was used to learn the model parameters and to perform the experiments reported in the next section. We reported the area under curve (AUC) and average error rate to evaluate the accuracy of the model. The error rate indicates the proportion of the cases in the test data for which the network predicted an incorrect value, where the prediction was taken as the state with highest belief.

$$error\ rate = \frac{\#false\ positive + \#false\ negative}{\#positive + \#negative}$$

The average error rate is reported based on cross-validation runs. The AUC and confidence interval were computed by cvAUC package [24] in R.

4. Results

4.1. Evaluation of outbreak detection methods

Table 2 summarizes the overall performance for 1800 outbreaks in terms of the minimum, maximum and average detection, specificity, and the median timeliness for each algorithm. Note, however, that there is considerable variability based on outbreak characteristics. The plot of performance evaluation is shown in Appendix A.

On average, W3 has the highest number of detected outbreaks (Mean of detection = 0.75), although the average specificity for this algorithm is low (mean specificity = 0.53). On average, the most reasonable performance belongs to C3 algorithm (mean specificity = 0.78, mean of detection = 0.72). The best timeliness is achieved by the W3 algorithm (0.83), at the expense of high rate false alarms. With this information in mind, we are interested to build a Bayesian network model to guide algorithm selection or parameter tuning for a surveillance application in order to improve the results, and to estimate the expected performance of an algorithm given a particular parameters setting or surveillance scenario.

4.2. Model evaluation

Among all BN structure learning methods examined, we found the structure learned by Taboo search resulted in the best prediction of the "detection" variable. This structure of the BN model learned from the data described in Table 1 is presented in Fig. 1.

From the network structure, it is clear that the mean of the baseline surveillance time series has a direct relationship with the timeliness of outbreak detection. It is also apparent that the baseline mean and variance are not independent, providing

[1] Technically, the lack of edges implies certain conditional independencies; for a detailed discussion, see Pearl's book [20] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. 1988. Morgan Kaufmann.

Table 2
Performance of a number of routinely used detection algorithms on simulated surveillance data.

Algorithm	Detection performance					
	Detection		Specificity		Timeliness	
	Min–max	Mean	Min–max	Mean	Min–max	Median
C1	[0, 1]	0.459	[0.377, 1]	0.813	[0, 1]	0.721
C2	[0, 1]	0.660	[0.013, 1]	0.765	[0, 1]	0.800
C3	[0, 1]	0.724	[0.013, 1]	0.785	[0, 1]	0.818
W2	[0, 1]	0.665	[0.082, 1]	0.754	[0.5, 1]	0.818
W3	[0, 1]	0.755	[0.054, 0.903]	0.533	[0.667, 1]	0.833
Poisson Regression	[0, 1]	0.658	[0.448, 1]	0.872	[0.496, 1]	0.786
Poisson Regression with guardband	[0, 1]	0.675	[0.446, 0.99]	0.870	[0.536, 1]	0.789

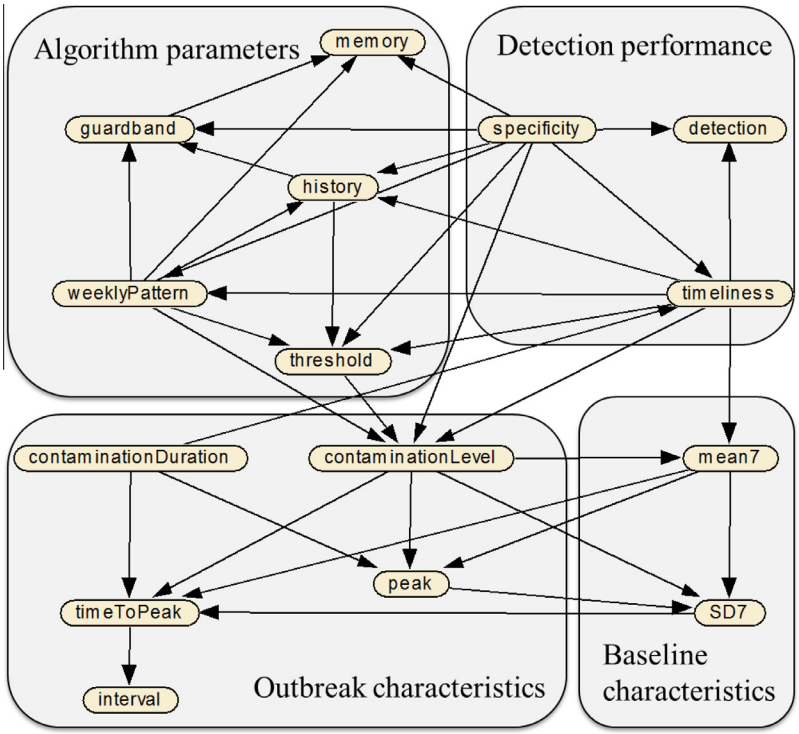

Fig. 1. The Bayesian network model learned from algorithm benchmarking data.

support for the choice of modeling such time series as Poisson stochastic processes. The graph structure also indicates that the dependency of algorithm parameters on baseline characteristics is mediated entirely through outbreak characteristics.

As expected, the relationship between outbreak characteristics and algorithm parameters is mediated almost entirely through the level of contamination and consequently the peak magnitude of the outbreak signal and the time from onset until the peak of the outbreak. These same outbreak characteristics also have a direct relationship with the timeliness and specificity of detection. The overall duration of the outbreak has only an indirect association with algorithm parameters and detection performance.

All algorithm parameters have a direct relationship with at least one metric of detection performance, but the threshold, history and adjustment for weekly patterns have direct relationships with both the specificity and the timeliness of outbreak detection. The use of a guardband and memory has a direct association with specificity, but only an indirect relationship with timeliness.

We evaluated the BN model presented in Fig. 1 in two scenarios, in order to assess its ability to predict the "detection" and "timeliness" variables simultaneously. Note that the goal of the BN model is not to detect the outbreaks, but rather to predict the performance of detection methods. The evaluation is based on 5-fold cross validation in which for each fold, 80% of the data were used for training the model and 20% were retained for testing. In the first scenario, we provided the information related to all algorithm parameters, baseline, and outbreak characteristics when predicting detection and timeliness. This scenario corresponds to the use of the model to explore determinants of outbreak detection with the goal of generating new insights or guiding the development of new algorithms. The ROC curve for the prediction of "detection" is presented in Fig. 2 and has an area under the curve (mean AUC) of 0.94 with a 95% confidence interval (0.91, 0.97). The error rates of predicting "detection" and "timeliness" were 10% and 24% respectively.

In the second evaluation scenario, we did not provide information related to the outbreak parameters, as in practice this information would generally not be available, or might be very imprecise. This scenario corresponds to the use of the model to identify and configure an algorithm for use in a public health setting. Fig. 3 shows the ROC curve for the prediction of "detection". In this case, the mean AUC was 0.86 with a 95% confidence interval of (0.83, 0.88) and the error rate was 20%. The presented ROC curves are based on the best folds in the cross-validation results, however

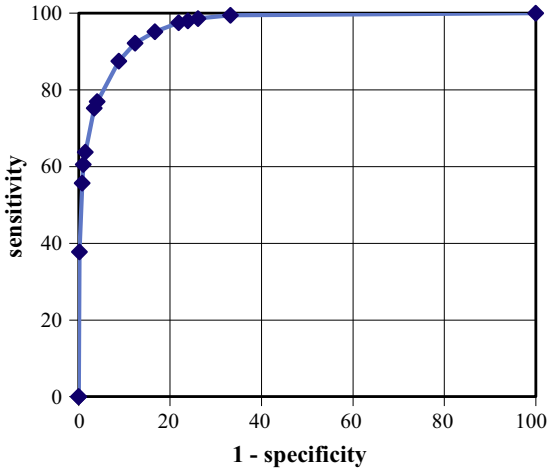

Fig. 2. The ROC curve of the BN model for prediction "detection" variable.

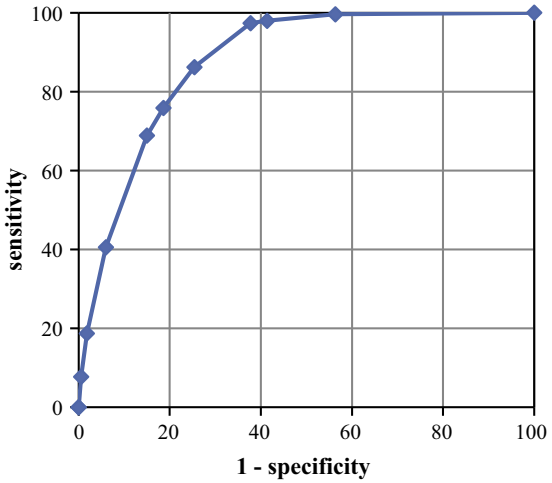

Fig. 3. The ROC curve for the BN model predicting "detection" when outbreak parameters are unknown.

the confidence intervals show that they were not be different from the general result.

4.3. Using the Bayesian network model

One of our objectives in developing the BN model is to guide the selection of detection algorithms in a surveillance application. To demonstrate the use of the model for this purpose, we apply the model in two scenarios (Table 3). The first column of Table 3 shows the variables that we set to a specific value or range, depending on the surveillance situation, including the expected outbreak specifications, tolerance for false alarms (which might depend on the resources available to further investigate alarms), or the threshold of sensitivity and timeliness of outbreak detection that is essential for a particular population need. The second column shows the probabilities or values of other variables inferred from Bayesian network model together with the probabilities of the "detection" and "timeliness" variables.

In the first scenario (first row of Table 3), we assume surveillance is focused on detection of a large GI outbreak (with contamination level of 10^{-4}) and can tolerate up to one false alarm every ten days (which corresponds to a specificity greater than 0.92). The inferred values for the algorithm parameters that result in the highest expected detection probability are listed in the second column: a threshold of less than 1.2 and a sliding window of seven days. The expectation of the detection probability, 0.99 and the resulting best timeliness, 0.78, are also presented in this row. These results are much better with respect to all performance metrics than what we would expect on average from all the different detection methods presented in Table 2. While, the assumption of high contamination makes all algorithms perform better than average, without the inference from the BN model it would have been difficult to estimate quantitatively the expected performance under different algorithm settings.

In the second scenario, we assume a smaller outbreak (with contamination level of 10^{-6}) and a specificity greater than 0.92. In this scenario, the best algorithm settings as listed in the second column: a threshold of less than 1.2, sliding window of seven days and no adjustment for weekly patterns. Following those settings, outbreaks with low contamination can still be detected with probability of 0.7 and timeliness of 0.54. These results are better than Table 2 with respect to the specificity, while similar to W2 and C2 algorithms in terms of sensitivity. From these scenarios, we can conclude that among outbreak detection algorithms examined, C2 and W2 have the highest sensitivity for large outbreaks when the desired specificity is greater than 0.92. Moreover, we can use the inference from the BN guide the configuration of these methods as the results in Table 2 suggest that a threshold below 1.2 will give the best detection performance in the examined scenarios.

We can also use the BN model to guide the configuration of different detection algorithms for a given scenario. For example, in order to detect outbreaks with high contamination using the C2 algorithm, the alerting threshold should be smaller than 1.2. The higher detection performance of the C1 algorithm is obtained when the alerting threshold is between 3.75 and 5. The Adaptive Poisson Regression algorithm is more sensitive to outbreaks when the alerting threshold is smaller than 2.4.

Another use of the Bayesian network model is to predict, with high accuracy, the performance of a detection method with a specific parameter setting. In Table 4, we show some example scenarios of this kind. The first row of this table assumes a detection method with adjustments for weekly pattern, sliding window of size seven, threshold between four and five, two days of guardband, and no memory. The expected detection performance in terms of all three metrics corresponding to specificity, detection probability, and timeliness are predicted to be 0.9, 0.63, and 0.52 respectively. The what-if scenario on the second row quantifies

Table 3
Using Bayesian network for inferential analysis in different surveillance situations.

What-if scenario	Inferred by Bayesian network model	
Specificity > 0.92 contamination level = 10^{-4} (reasonable specificity for high contamination outbreaks)	Algorithm setting with highest detection probability: 0 < threshold < 1.2 history = 7 days	Expected detection prob.= 0.99 Expected timeliness = 0.78
Specificity > 0.92 contamination level = 10^{-6} (reasonable specificity for low contamination outbreaks)	Algorithm setting with highest detection probability: 0 < threshold < 1.2 history = 7 days no weekly patterns	Expected detection prob.= 0.7 Expected timeliness = 0.54

Table 4
Using Bayesian network for inferential analysis for different algorithm configurations.

What-if scenario	Inferred by Bayesian network model
Weekly pattern = 1 History = 7 4 < threshold < 5 Guardband = 2 No memory	Expected specificity = 0.9 Expected detection prob. = 0.63 Expected timeliness = 0.52
Weekly pattern = 1 History = 56 4 < threshold < 5 No guardband No memory	Expected specificity = 0.72 Expected detection prob. = 0.81 Expected timeliness = 0.58

the performance trade-off in sensitivity and specificity if we drop the guardband and increase the size of the sliding window from the scenario 1. The specificity decreases to 0.72 while the detection probability increases to 0.81, with a slight change in timeliness.

5. Discussion and conclusion

In this study, we analyzed outbreak detection performance for a range of algorithms that are commonly used in public health practice, considering a range of features related to the outbreak characteristics, baseline data, and the parameter settings for the detection methods. We assessed the performance of seven different outbreak detection algorithms using simulated and real surveillance data for GI outbreaks in eighteen outbreak scenarios and trained Bayesian networks to model the relationships between all surveillance attributes and the detection performance. Our evaluation results show that even when the outbreak characteristics were unknown *a priori*, the model was able to predict detection performance with high accuracy (AUC = 0.86).

The Bayesian network model developed in this paper allows quantifying the effect of outbreak characteristics and algorithm configurations on the performance of detection algorithms. As expected, the most informative determinants of detection performance were the alerting threshold, which is a parameter of the detection method, and the contamination level and the peak size of the outbreak. But our model also quantified the contribution of other algorithm features such as accounting for day-of-week and maintaining a guardband or memory. We demonstrated how inference performed using our model can help to develop what-if analyses for using detection methods in practice, or to find an appropriate algorithm configuration given the desired level of detection performance for outbreak scenarios. Such an inferential tool gives insight into the features of detection methods that are important to provide better performance. We also described how the model can be utilized to predict the expected performance of detection methods in different surveillance situations.

One limitation in comparing surveillance methods is the lack of data for benchmarking. This limitation was addressed in our work by using simulated data. Our approach is similar in spirit to the research reported by Lewis and colleagues [25], who used simulation of influenza outbreaks to evaluate spatiotemporal outbreak detection methods. However, they did not quantify the effects of algorithm parameters on detection performance, and we believe this is an important contribution of our research.

Our approach can be extended to allow a coherent evaluation of new algorithms and new data sources as needed. In particular, using our current model, we can evaluate outbreak detection performance for new algorithms different than the C, W, and Adaptive Poisson algorithms. Any configuration of considered parameters in our model different than the ones belonging to these algorithms can be thought of as a new detection method and can also be evaluated.

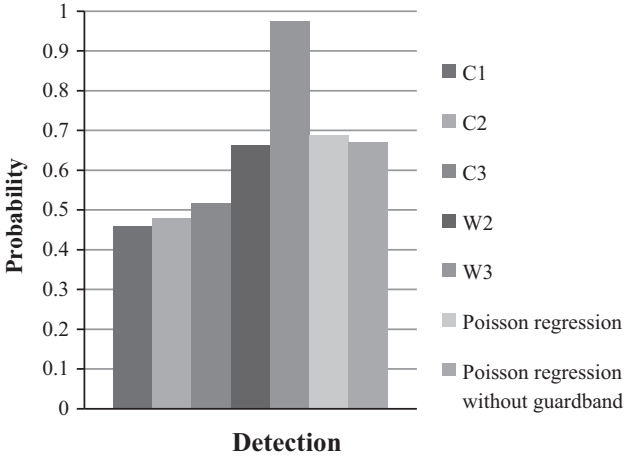

Fig. A.1. Detection rate of a number of routinely used detection algorithms on simulated surveillance data.

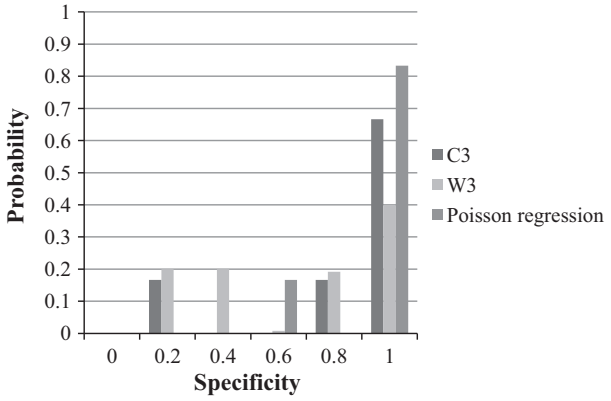

Fig. A.2. Specificity of three detection algorithms on simulated surveillance data.

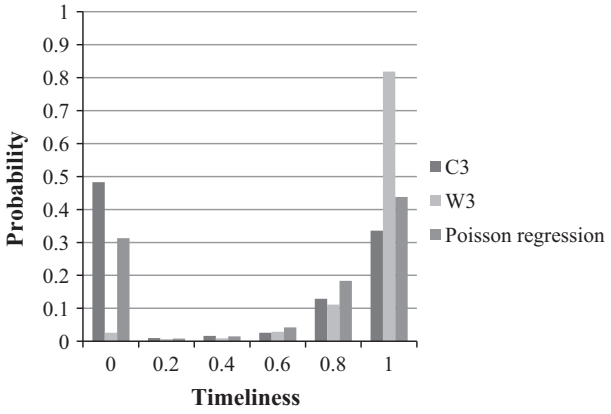

Fig. A.3. Timeliness of three detection algorithms on simulated surveillance data.

We presented several scenarios of outbreaks and desired performance, and used inference to suggest the best algorithm and parameter setting to use, as well as to quantify the expected performance. These scenarios are by no means exhaustive, and are meant as examples of what kinds of inference can be performed. Of course, in order to use this model as a tool for what-if analysis in the public health sector, an adequate interface would also need to be developed, but this goes beyond the scope of our work.

We used emergency department visits as the baseline time series for outbreak detection. In recent years, non-traditional data sources have been introduced in public health and surveillance systems. This includes mobile phone data [26], social data [27], micro-blogging [28], Twitter feeds and Google search queries [29]. While introducing and combining new data sources, especially in the era of big data, are promising directions for research in biosurveillance systems, the evaluation of their relevance and significance will be extremely important. Evaluation studies such as [30] are needed to compare these new data sources to the existing ones, and the approach that we describe could be used to consider the relative contribution to detection performance of data sources and algorithms.

A number of extensions to this work may improve the generalizability of the results. We used simulated outbreaks superimposed on real surveillance data; therefore, the results are affected by the quality of the simulation. Our approach can be extended by including more detection methods, using spatio-temporal data simulations, and considering health care utilization data sources in addition to ED visits.

Funding statement

This work was funded by the following CIHR grants: Evaluating Syndromic Surveillance in Public Health Practice (MOP-84493), Evidence-Based Algorithm Selection in Public Health Surveillance (MOP-93587).

Contributorship statement

All authors listed in the paper provided substantial contribution in all stages of conception, design, analysis and interpretation of the results, drafting the manuscript or revising it critically.

Acknowledgments

We would like to thank the Canadian Institutes of Health Research (CIHR) for providing the funding for this research. We thank Anya Okhmatovskaia for her help in generating the simulation data used in this paper.

Appendix A

Fig. A.1 shows the detection probability of different detection algorithms for 1800 outbreaks respectively.

Figs. A.2 and A.3 show the probability of different values of specificity and timeliness of three algorithms for the data set respectively. The performance of other algorithms is very close to the illustrated ones of the same family.

References

[1] Jackson ML, Baer A, Painter I, Duchin J. A simulation study comparing aberration detection algorithms for syndromic surveillance. BMC Med Inform Decis Mak 2007;7:6.
[2] Watkins RE, Eagleson S, Hall RG, Dailey L, Plant AJ. Approaches to the evaluation of outbreak detection methods. BMC Public Health 2006;6:263.
[3] Buckeridge DL, Okhmatovskaia A, Tu S, O'Connor M, Nyulas C, Musen MA. Predicting outbreak detection in public health surveillance: quantitative analysis to enable evidence-based method selection. In: AMIA annual symposium proceedings: American medical informatics association; 2008. p. 76.
[4] Izadi M, Buckeridge D, Okhmatovskaia A, Tu SW, O'Connor MJ, Nyulas C, et al. A Bayesian network model for analysis of detection performance in surveillance systems. In: AMIA annual symposium proceedings: American medical informatics association; 2009. p. 276.
[5] Hutwagner ML, Thompson MW, Seeman GM, Treadwell T. The bioterrorism preparedness and response early aberration reporting system (EARS). J Urban Health 2003;80:i89–96.
[6] Tokars JI, Burkom H, Xing J, English R, Bloom S, Cox K, et al. Enhancing time-series detection algorithms for automated biosurveillance. Emerg Infect Dis 2009;15:533.
[7] Xing J, Burkom H, Tokars J. Method selection and adaptation for distributed monitoring of infectious diseases for syndromic surveillance. J Biomed Inform 2011;44:1093–101.
[8] Shewhart Walter A. Statistical Method from the Viewpoint of Quality Control. Mineola, NY: Dover Publications; 1939.
[9] Page E. Continuous inspection schemes. Biometrika 1954:100–15.
[10] Douglas CM. Introduction to Statistical Quality Control. John Wiley & Sons; 2005.
[11] Shiryaev AN. On optimum methods in quickest detection problems. Theor Probab Appl 1963;8:22–46.
[12] Frisén M, De Maré J. Optimal surveillance. Biometrika 1991;78:271–80.
[13] Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. J Roy Stat Soc: Ser A (Stat Soc) 2003;166:5–21.
[14] Lu H-M, Zeng D, Chen H. Prospective infectious disease outbreak detection using Markov switching models. In: Knowledge and data engineering, IEEE transactions on, vol. 22; 2010. p. 565–77.
[15] Lombardo JS, Buckeridge DL. Disease surveillance: a public health informatics approach. Wiley-Blackwell; 2007.
[16] Okhmatovskaia A, Verma AD, Barbeau B, Carriere A, Pasquet R, Buckeridge DL. A simulation model of waterborne gastro-intestinal disease outbreaks: description and initial evaluation. In: AMIA annual symposium proceedings: American medical informatics association; 2010. p. 557.
[17] Buckeridge DL, Jauvin C, Okhmatovskaia A, Verma AD. Simulation analysis platform (SnAP): a tool for evaluation of public health surveillance and disease control strategies. American Medical Informatics Association; 2011. p. 161.
[18] McKay MD, Beckman RJ, Conover WJ. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 1979;21:239–45.
[19] Netica Bayesian network software from Norsys.
[20] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann; 1988.
[21] Pearl J. Causality: models, reasoning and inference. Cambridge Univ Press; 2000.
[22] Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley; 1973.
[23] Koller D, Friedman N, Getoor L, Taskar B. 2 Graphical models in a Nutshell. Stat Relat Learn 2007:13.
[24] LeDell E, Petersen M, van der Laan M, LeDell ME. Package 'cvAUC'.
[25] Lewis B, Eubank S, Abrams AM, Kleinman KP. In silico surveillance: evaluating outbreak detection with simulation models. BMC Med Inf Decis Mak 2013;13:12.
[26] Buckee CO, Wesolowski A, Eagle NN, Hansen E, Snow RW. Mobile phones and malaria: modeling human and parasite travel. Travel Med Infect Dis 2013.
[27] Alasaad S. War diseases revealed by the social media: massive leishmaniasis outbreak in the Syrian Spring. Parasites Vectors 2013;6:94.
[28] Donelle L, Booth R. Health tweets: an exploration of health promotion on twitter. Online J Issues Nurs 2012;17:4.
[29] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2008;457:1012–4.
[30] Wilson N, Mason K, Tobias M, Peacey M, Huang Q, Baker M. Interpreting Google flu trends data for pandemic H1N1 influenza: the New Zealand experience. Euro Surv: Bull Eur Mal Trans – Eur Commun Dis Bull 2009:14.