A Bayesian Network Model for Analysis of Detection Performance in Surveillance Systems

Masoumeh Izadi[1], David Buckeridge[1], Anna Okhmatovskaia[1],
Samson W. Tu[2], Martin J. O‘Connor[2], Csongor Nyulas[2], Mark A. Musen[2]
[1]McGill University, Montreal, QC; [2]Stanford University, Palo Alto, CA

Abstract

Worldwide developments concerning infectious diseases and bioterrorism are driving forces for improving aberrancy detection in public health surveillance. The performance of an aberrancy detection algorithm can be measured in terms of sensitivity, specificity and timeliness. However, these metrics are probabilistically dependent variables and there is always a trade-off between them. This situation raises the question of how to quantify this trade-off. The answer to this question depends on the characteristics of the specific disease under surveillance, the characteristics of data used for surveillance, and the algorithmic properties of detection methods. In practice, the evidence describing the relative performance of different algorithms remains fragmented and mainly qualitative. In this paper, we consider the development and evaluation of a Bayesian network framework for analysis of performance measures of aberrancy detection algorithms. This framework enables principled comparison of algorithms and identification of suitable algorithms for use in specific public health surveillance settings.

Introduction

Outbreaks of infectious diseases occur regularly and result in substantial cost and morbidity [10]. Unfortunately, the risk of future outbreaks is considerable due to the continuing emergence of new diseases and the limitations of our current systems [5, 14]. If future outbreaks are detected rapidly, however, effective interventions exist to limit the health and economic impacts [4, 17]. Traditional public health surveillance systems are expected to detect disease outbreaks, but these systems have failed to detect many such outbreaks, including the SARS outbreak in Toronto, the Cryptosporidiosis outbreak in Milwaukee, and the E. coli outbreak in Walkerton. These failures had tragic consequences, including thousands infected and many deaths [15, 12, 16]. Reviews of the public health response following these and other outbreaks consistently call for improvements to the public health surveillance infrastructure. In response, many public health agencies have adopted syndromic surveillance systems, which acquire data in real-time from clinical and other settings, group records into broad syndromes, and apply statistical algorithms to detect aberrancies. Many aberrancy detection algorithms have been introduced in the last decade [7, 9]. However, these algorithms perform differently when applied to different data sets in different situations [2]. Evidence describing the performance of these algorithms under various conditions remains limited and mainly qualitative [1]. It is important to be able to select an algorithm, with a particular parameter tuning in a particular surveillance application, with good level of confidence on its performance.

In our earlier work [3], a model of surveillance data and outbreak signals was created. We used BioSTORM [13] as a testbed to evaluate algorithms used widely by the surveillance community and to assess the accuracy and timeliness of these algorithms under different parameter settings; the results of these evaluation studies were used to create a database; and a logistic regression model was used to predict the ability of different algorithms to detect different types of outbreaks in several surveillance configurations using this database. While the work generates insights, we noted limitations of logistic regressions in handling multiple outcomes in a single model and in allowing for complex relationships between covariates. In this paper, we address these limitations by developing a framework for reasoning under uncertainty about the performance of outbreak detection algorithms. This framework permits a more flexible representation of dependencies between the variables involved and represents different performance metrics in a single model. This representation is essential for quantifying the trade-offs between performance measures. In addition to predicting algorithm performance in a unified form, our model allows us to discover knowledge about the performance of aberrancy detection algorithms used in public health surveillance.

Method

We used Bayesian networks in probabilistic evaluation of detection methods to answer the question of which algorithmic setting is more likely to result in a desirable overall performance. A Bayesian network [8] is a directed acyclic graph (DAG), in which nodes represent random variables and edges represent conditional dependencies between variables. Each nodes is associated with a conditional probability table (CPT). The probability of a node can be calculated when the values of its incoming nodes are known. To describe a Bayesian network we need to specify the graph structure and the values of each CPT based on data. Conceptually, a Bayesian network can help to answer questions about the features of algorithms that are important in different surveillance contexts. For instance, we might be interested to set as evidence the types of outbreaks expected, and ask which algorithmic features will maximize sensitivity while observing the concurrent effect on timeliness. For the purpose of demonstration in this paper, we selected the same set of algorithms, data signals, and outbreak characteristics that we used in our earlier work [3] to allow a direct comparison between the Bayesian network results and our earlier results from logistic regression analysis.

In order to collect the evidence about the detection performance of the selected methods, we ran different surveillance scenarios on a variety of detection configurations and stored the results. The configurations and the results will serve as input to Bayesian network model construction. In this section we briefly describe this input data and their corresponding variables in the network.

Baseline Surveillance Data

We used a set of surveillance data created by researchers at the CDC [6] for the objective of comparison between detection algorithms. Each data set contains 1000 simulated daily time series over 6 years (1994-1999) containing no outbreaks. The features of these data include resolution, seasonality, trend, mean, and variance. Means and standard deviations in the baseline data were based on observed values from national and local public health systems and syndromic surveillance systems, with the adjustments made for days of the week, holidays, post-holiday periods, seasonality, and trend [6].

Outbreak Characteristics

A variety of simulated outbreak signals, also developed by researchers at the CDC, were superimposed on the baseline data. The signals included 1-day spikes and multi-day signals generated using lognormal or inverse lognormal functions. We characterized these signals using non-parametric measures of the signal, such as the peak size expressed in the units of standard deviation of the baseline and the day of the peak amplitude of the signal. Table 1 present these variables and their discretization level used in our Bayesian network model.

Detection Methods

We choose a family of algorithms widely used in public health surveillance, C1, C2 and C3 algorithms used in the Early Aberration Reporting System (EARS) software developed by the CDC [6]. These algorithms make an alerting decision for each day in a time series by comparing a test statistic to a historical mean calculated from the seven previous days in the baseline data and are closely related to statistical process control methods such as Shewhart charts [11] and cumulative sums (CUSUMs). More precisely, C-algorithms are defined as variants of single-sided CUSUM. The computation of a test statistic for these algorithms is different from traditional cumulative summation in that C1 and C2 use only the current observation, which makes them more similar to Shewhart charts and the C3 algorithm sums two previous observations and the current one. A true cumulative sum, on the other hand, can be influenced by an infinite number of prior observations. The methods C1, C2, and C3 were named according to their degree of sensitivity, with C1 being the least sensitive and C3 the most sensitive. The distinctions among the three C-algorithms are in a two-day guardband [1] and the inclusion of two recent observations (memory) in the computation of a test statistic. C1 uses neither guard-band nor memory; C2 uses the guard-band, but not memory; and C3 uses both. All three algorithms were also configured to use a range of alerting thresholds corresponding to varying specificity levels. Table 2 shows the features related to these algorithms.

Database Specification

In specifying the performance measures, we ran a large number of experiments using the BioSTORM (Biological Spatio-Temporal Outbreak Reasoning Module) software [13]. Each experiment assigned a different set of parameters values to the outbreaks and detection algorithms. The results for detection and timeliness as well as the settings for algorithm parameters, outbreak parameters, and data sources were recorded in a database. We applied each of the C algorithms at 10 different levels of specificity to each signal producing a total of $1,076,880$ observations of the variables listed in Table 2. In $748,964$ (69.5%) of observations, an outbreak was detected.

Model Specification

We use a Bayesian network to quantify relationships between features of data types, outbreak signals, detection

[1] Guard-band is a buffer period or a time interval between the baseline and test periods. Guard-band is useful to separate the observed values from the data used to calculate the historical mean in the baseline, in particular when early undetected outbreak effects may inflate the baseline data and increase the data expectation

Table 1. Outbreak parameters

Variable	Description	Range	Levels
Duration	Length of outbreak signal in days	0-16	0-1, 1-8, 8-16
Peak day	Day when peak signal occurs within outbreak	0-15	0-4, 4-15
Peak size	Number of standard deviations of outbreak signal above baseline mean	binary	2,3
Spike	Outbreak is a single-day spike	binary	0,1

Table 2. Algorithm parameters

Variable	Description	Range	Levels
Specificity	Proportion of non-outbreak days with alarm	0.87-0.97	0.87-0.89, 0.89-0.91, 0.91-0.93, 0.93-0.95, 0.95-0.97
Guardband	Guardband of two days is used	binary	0,1
Memory	Memory of past observations is used to compute running sum	binary	0,1
Detected	Sensitivity per outbreak	binary	0,1
Detection day	Day when outbreak detected	0-15	0-2, 2-15

methods, and outbreak detection performance that exist in the database we generated through the experimental studies explained above. We initially found the skeleton structure of the network from the data with no missing values using the structure learning method of Max-Min Hill-Climbing (MMHC) [18]. This type of search is very useful when dealing with large data sets like ours because of its computational efficiency. We refined the structure manually to incorporate the domain knowledge about associations among the variables. More precisely, we reversed the direction of the edges based on the Markov equivalence property where possible, in order to better model the known causal relationships. The model obtained is shown in Figure 1. In the network,we represented a spike as simply an outbreak of 1-day duration, which is encoded as a level of the duration variable. Prior to observing the data, a uniform distribution of probabilities among all states of a node was assigned to each CPT. These probabilities were then updated through a learning process, as we provide each data instance. The CPTs of our network in Figure 1 were learned from our database explained earlier, using expectation maximization (EM)algorithm in NeticaTM 2.0 software. EM learning repeatedly tries to find a better network (in log likelihood sense) by doing an expectation (E) step followed by a maximization (M) step. Therefore, the log likelihood of the new net is always as good as or better than the previous one. This process is repeated until the log likelihood numbers are no longer improving enough, or the desired number of iterations has been reached. These quantities can be specified by the user. Given the structure of a BN, EM converges to an optimal solution for the parameters, according to the data.

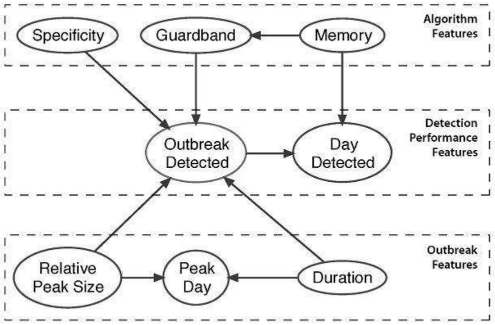

Figure 1. A representative section of the Bayesian network model for aberrancy detection

Model Evaluation and Applications

We evaluated the accuracy of the Bayesian network model for predicting sensitivity and timeliness. A 10-fold cross validation was used to evaluate the predictive ability of the Bayesian network. We used the network model for different analyses by providing evidence in some nodes and infer the values for others. Evidence is provided about variables by manually setting the probability for the value of a feature to 1 or 0. The number of cases in the database that support an observation is recorded in order to provide the confidence interval for the posterior belief. The initial belief for the value of each variable is the marginal probability for each node after estimating the conditional probabilities for the network. We particularly used the model to infer how different algorithms and outbreak characteristics could influence algorithm performance. The model allows for three-way reasoning and quantization of the trade-offs between sensitivity, specificity and timeliness. For example, we can provide the specificity measures as evidence and infer the changes in sensitivity and timeliness. Using the Bayesian network that we constructed for the EARS algorithms, we set as evidence the types of outbreaks expected, and inferred which algorithm features are likely to increase sensitivity while observing the concurrent effect on timeliness.

Results

In our model evaluation we found that the accuracy of Bayesian models in predicting whether or not an outbreak would be detected (variable Detected, i.e. sensitivity) by C-algorithms on detection of multi-day outbreaks had an area under the curve of 0.79, similar to the logistic regression results [3]. The Bayesian network model had a smaller error rate of 18% for predicting the Detected variable and an error rate of 22% for Detection Day given the specified resolution for these variables as in Table 2.

Table 3 summarizes the results of inference analysis where the outbreak was assumed to have a peak size of three times the standard deviation of the baseline data, the duration of the outbreak was between 8 and 16 days, and the peak of the outbreak occurred before the fourth day. In the three last columns, we infer from the network sensitivity and timeliness for the three EARS algorithms for a specific type of outbreak. We specify the algorithm by providing evidence about the value of algorithm variables (e.g., for C1, which has neither a guard band nor memory, we set both P(Guard band=True) and P(Memory=True) to 0) and the type of outbreak by providing evidence about the outbreak variables. The posterior belief is then estimated for the variables of interest, namely outbreak detected and day detected. Averaging over the full range of specificity examined in the experimental studies $(0.87-0.97)$, the results in this example indicate that inclusion of a guard band interval (i.e., moving from C1 to C2) increases sensitivity from 0.68 to 0.82 and improves the probability of detection in less than two days from 0.63 to 0.70. Inclusion of a memory parameter improves sensitivity further, from 0.82 to 0.87 and improves the probability of detection within two days slightly, from 0.70 to 0.72. All of these differences are statistically significant (standard error of each point estimate is very small due to the large number of observations and the relatively simple structure of the network).

The effect of specificity on time to detection was investigated through providing evidence on distinctive algorithmic features among C-algorithms (Memory and Guard band). The results are presented in Figure 2. It is apparent that different levels of specificity have different levels of impact on these algorithms. This difference is more pronounced as the false alarm rate decreases with higher levels of specificity. We performed this analysis without providing any evidence on the sensitivity of these algorithm, although information about the sensitivity can be inferred at the same time. It should be noted that sensitivity can be treated as a special case of timeliness in which the detection time is infinite. Therefore we can join the two corresponding nodes in the network, if the accumulated sensitivity over time is not important.

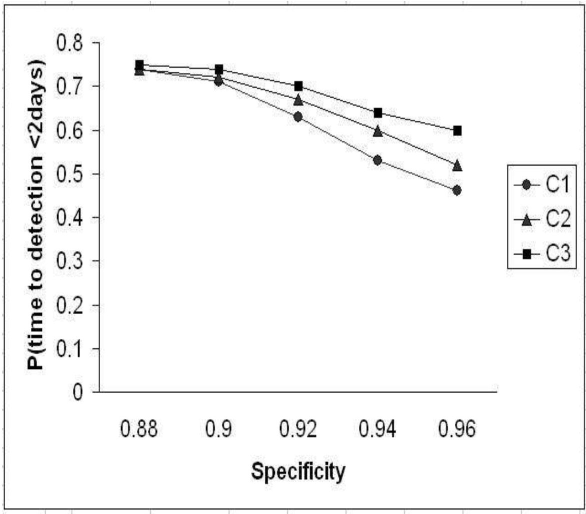

Figure 2. The relationship between the specificity of a detection method and likelihood of detecting an outbreak in less than two days by C-algorithms.

Discussion

Our results suggest that guard band inclusion leads to better performance in terms of sensitivity and timeliness. The guidance for a practitioner, therefore, is that inclusion of a guard band is important for detecting the type of outbreaks we considered and that also using a memory parameter will further improve detection performance. These results are very important given the widespread use of the EARS algorithms in public health surveillance. Another practical implication of our findings relates to resources used by public health institutions to handle false alarms. Results from Figure 2 show that the distinction between different C-algorithms is mainly on high specificities. Therefore, if high sensitivity is required and personnel and resources are available to handle high false alarms it does not matter much which algorithm to choose but if the resources are limited then algorithm C3 would be more desirable in providing the best timeliness among C-algorithms.

We studied limited number of features of outbreaks and a small number of algorithms. In the future, we plan to extend this framework as we encode additional algorithms in terms of our unified model and run additional experimental studies through BioSTORM and extend the database results.

Table 3. Example of inferences from the Bayesian network in Figure 1.

Feature Type	Variable	Value	Initial Belief	C1	C2	C3
Algorithm	Guard band	P(Guard band=1)	0.67	0.00	1.00	1.00
	Memory	P(Memory=True)	0.33	0.00	0.00	1.00
Detection	Detected	P(Detected=True)	0.70	**0.68**	**0.82**	**0.87**
	Day Detected	P(Day Detected¡ 2)	0.64	**0.63**	**0.70**	**0.72**
Outbreak	Peak Size	P(Peak Size=3)	0.50	1.00	1.00	1.00
	Peak Day	P(Peak Day $<$ 4)	0.69	1.00	1.00	1.00
	Duration	P(8 $<$ Duration $<$ 16)	0.40	1.00	1.00	1.00

We will iteratively re-learn the network structure and conditional probability tables, and use the network to update the determinants of outbreak detection and guide the use of outbreak detection algorithms.

Conclusion

There is currently very little explicit guidance available for public health practitioners as they attempt to select algorithms and tune them for use in syndromic surveillance systems. In this paper, we used a Bayesian network framework for reasoning under uncertainty about the performance of outbreak detection algorithms and discussed the development and evaluation of this model. We have already noted with our initial network for the C algorithms that although a guardband facilitates detection, it can also lead to later detection for certain types of outbreaks. Balancing this tradeoff between sensitivity and timeliness is of fundamental importance to surveillance and deserves further attention. Our promising results suggest further directions for research, including consideration of different types of outbreaks, wider range of algorithms and data sources.

Acknowledgments

This research was supported by a grant from Centers for Disease Control and Prevention under the BioSense Initiative to Improve Early Event Detection (RFA-PH-05-126) and Funds from Canadian Institute for Health Research, CIHR. Masoumeh Izadi is partially supported by McLaughlin fellowship and David Buckeridge is supported by Canada Research Chair Grant in Public Health Informatics.

References

[1] D. Bravata, K. McDonald, and W. Smith. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. *Ann Intern Med*, (140):910–922, 2004.

[2] D. Buckeridge. Outbreak detection through automated surveillance: A review of the determinants of detection. *Journal of Biomedical Informatics*, 40(3):370–379, 2007.

[3] D. Buckeridge, A. Okhmatovskaia, S.Tu, M. O‘Connor, C. Nyulas, and M. Musen. Predicting outbreak detection in public health surveillance: Quantitative analysis to enable evidence-based method selection. In *American Medical Informatics Association*, 2008.

[4] A. Fiore. Prevention and control of influenza: recommendations of the advisory committee on immunization practices (acip), 2008. *MMWR Recomm Rep*, 57(7):1–60, 2008.

[5] S. Hrudey and E. Hrudey. Walkerton and north battleford– key lessons for public health professionals. *Can J Public Health*, 93(5):332–3, 2002.

[6] L. Hutwagner, T. Browne, M. Seeman, and A. Fleischauer. Comparing aberration detection methods with simulated data. *Emerg Infect Dis*, (11):314–6, 2005.

[7] L. Hutwagner, E. Maloney, N. Bean, L. Slutsker, and S. Martin. Using laboratory-based surveillance data for prevention: an algorithm for detecting salmonella outbreaks. *Emerg Infect Dis*, 3:395–400, 1997.

[8] F. Jensen. *An introduction to bayesian networks*. Springer, 1996.

[9] M. Kulldorff. A spatial scan statistic. *Commun Stat Theory Methods*, (26):1481–96, 1997.

[10] J. Liang. Surveillance for waterborne disease and outbreaks associated with drinking water and water not intended for drinking: United states, 2003-2004. *MMWR Surveill Summ*, 55(12):31–65, 2006.

[11] D. Montgomery. *Introduction to Statistical Quality Control*. 1997.

[12] M.Wagner, V. Dato, J. Dowling, and M. Allswede. Representative threats for research in public health surveillance. *J. of Biomedical Informatics*, 36(3):177–188, 2003.

[13] C. Nyulas, M. O‘Connor, S. Tu, D. Buckeridge, A. Okhmatovskaia, and M. Musen. An ontology-driven framework for deploying jade agent systems. In *IEEE International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2008.

[14] P. Payment. Occurrence of pathogenic microorganisms in the saint lawrence river (canada) and comparison of health risks for populations using it as their source of drinking water. *Can J Microbiol*, 46(6):565–76, 2000.

[15] SARS. The sars commission, final report: Toronto, 2006.

[16] R. Stirling. Waterborne cryptosporidiosis outbreak, north battleford, saskatchewan, spring 2001. *Can Commun Dis Rep*, 27(22):185–92, 2001.

[17] A. Tricco. A review of interventions triggered by hepatitis a infected food-handlers in canada. *BMC Health Serv Res*, 6:157, 2006.

[18] I. Tsamardinos, L. Brown, and C. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.