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.