Journal of Biomedical Informatics 44 (2011) 1093–1101

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Method selection and adaptation for distributed monitoring of infectious diseases for syndromic surveillance

Jian Xing [a,*], Howard Burkom [b], Jerome Tokars [a]

[a] Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta, GA 30333, USA
[b] The Johns Hopkins University, Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723, USA

ARTICLE INFO

Article history:
Received 7 April 2010
Accepted 18 August 2011
Available online 24 August 2011

Keywords:
Automated disease surveillance
Outbreak detection
Algorithm evaluation
Poisson regression

ABSTRACT

Background: Automated surveillance systems require statistical methods to recognize increases in visit counts that might indicate an outbreak. In prior work we presented methods to enhance the sensitivity of C2, a commonly used time series method. In this study, we compared the enhanced C2 method with five regression models.

Methods: We used emergency department chief complaint data from US CDC BioSense surveillance system, aggregated by city (total of 206 hospitals, 16 cities) during 5/2008–4/2009. Data for six syndromes (asthma, gastrointestinal, nausea and vomiting, rash, respiratory, and influenza-like illness) was used and was stratified by mean count (1–19, 20–49, $\geqslant$50 per day) into 14 syndrome-count categories. We compared the sensitivity for detecting single-day artificially-added increases in syndrome counts. Four modifications of the C2 time series method, and five regression models (two linear and three Poisson), were tested. A constant alert rate of 1% was used for all methods.

Results: Among the regression models tested, we found that a Poisson model controlling for the logarithm of total visits (i.e., visits both meeting and not meeting a syndrome definition), day of week, and 14-day time period was best. Among 14 syndrome-count categories, time series and regression methods produced approximately the same sensitivity (<5% difference) in 6; in six categories, the regression method had higher sensitivity (range 6–14% improvement), and in two categories the time series method had higher sensitivity.

Discussion: When automated data are aggregated to the city level, a Poisson regression model that controls for total visits produces the best overall sensitivity for detecting artificially added visit counts. This improvement was achieved without increasing the alert rate, which was held constant at 1% for all methods. These findings will improve our ability to detect outbreaks in automated surveillance system data.

Published by Elsevier Inc.

1. Introduction

Automated surveillance involves using algorithms to monitor pre-existing datasets for evidence of disease outbreaks or trends. Administrative data, such as the free text field recording each patient's chief complaint for the visit, are commonly used. These chief complaint strings are used to classify the records according to a set of disease types or syndrome groupings, denoted simply as *syndromes* in the discussion to follow. For example, a record with a chief complaint of "deep cough with chills" could be classified in the respiratory syndrome. For each syndrome, daily counts of records are stored and combined to form syndromic time series, i.e., syndrome counts plotted over time. These time series are monitored using automated statistical algorithms for increases that are anomalous compared to expected or modeled behavior [1–4]. Typically, monitoring for increased counts on a given index day involves calculating the expected number of visits on the day, assuming no outbreak, and comparing this expected value to the count actually observed.

The measures customarily used in epidemiology, incidence rates of disease in the population at risk, are not available in automated surveillance. Instead, automated surveillance aims to measure changes in healthcare behavior, such as visits to an emergency department. Since a true denominator is not available to calculate disease incidence, simple counts (numerator data) are often monitored. An alternative is to use as a surrogate rate the proportion of total visits that are classified into a given syndrome. The "C2-rate" method [5] is an implementation of this approach. This method involves calculating the expected number of visits for a given syndrome on an index day as follows: summing the visits for the syndrome over the recent 2–4 weeks, summing total visits over

* Corresponding author. Address: 1600 Clifton Road NE, MS G-37, Atlanta, GA 30333, USA. Fax: +1 404 718 8585.
E-mail addresses: jxing@cdc.gov, esw4@cdc.gov (J. Xing), Howard.Burkom@jhuapl.edu (H. Burkom), jit1@cdc.gov (J. Tokars).

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2011.08.012

the same time period, calculating the proportion of total visits that were for the syndrome, and multiplying this proportion by the total number of visits on the index day. This method was shown to produce more accurate expected values and better sensitivity for detecting an increased number of visits than basing expected values on the simple mean number of recent visits for the syndrome.

Our previous work on comparison of regression models to control charts, using syndromic ED time series collected by the BioSense surveillance system [6] operated by the US Centers for Disease Control and Prevention (CDC), was reported [7]. Except for the broadest syndrome groupings at large facilities, daily syndromic time series at the facility level are often sparse, with many daily counts of zero. The rationale for using regression modeling for surveillance is to capture systematic data effects such as trends or cycles so that algorithms applied to forecast residuals are not biased by these effects [8]. However, the sparseness of day-of-week and seasonal effects at the facility level does not work well for regression [7]. The widely used Serfling models are generally applied at the city level or higher [9]. Therefore, the series examined in this manuscript were restricted to the level of the Metropolitan Reporting Area (MRA) including visits from multiple facilities. The regression variables described in the methods were based on earlier findings with the same data [7]. The current study includes expanded methods applied to a more extensive set of syndromic time series with more rigorous statistical examination.

In this work we applied alerting algorithms to regionally distributed time series and report findings from a biosurveillance model comparison study. The objective of the study was to select the most appropriate statistical alerting algorithms for monitoring of disparate surveillance data types at distributed sites. This distributed monitoring should be robust in that (a) forecast and anomaly detection methods are appropriate for the various time series of interest, and (b) the detection performance, measured by sensitivity and background alert rate, is consistent across data types and monitoring sites.

2. Methods

2.1. Description of study data

Study data were taken from the CDC BioSense automated biosurveillance system. This system includes emergency department records with a free text field for the patient's chief complaint during the visit. We chose hospital Emergency Department free-text chief complaint data sent by 258 non-federal hospitals nationwide. Each hospital belongs to one of 48 Metropolitan Reporting Areas (MRA). Each MRA includes between 1 and 39 hospitals. For our study we chose 16 MRAs such that the data reported by each MRA included at least five hospitals (total of 206 hospitals) and such that the proportion of unreported historical data days was <1%. To study model forecast and detection performance over a variety of data scales, seasonal behaviors, and regions, we selected six BioSense syndromes and sub-syndromes [6,16]. The Gastrointestinal syndrome and the Nausea and Vomiting sub-syndrome were reported in high counts from most hospitals. The Respiratory syndrome and Influenza-like illness (ILI) sub-syndrome show a seasonal pattern. The Asthma sub-syndrome, a subset of the Respiratory syndrome, was reported in low counts from most hospitals. Finally, the Rash syndrome has intermediate visit counts.

The study period was 5/1/2008–4/30/2009. We used data from a 56-day baseline period to compute regression coefficients. Thus the baseline period for the first test day began on 3/4/2008. For coefficients reflecting recent data behavior and including the data from hospitals providing data in recent weeks, we used a sliding baseline, so that each baseline period ended 2 days before the date whose data was tested for an anomaly. The purpose of the 2-day buffer was to avoid contamination of the baseline data by the early phases of an outbreak. The data for this study were provided by 16 Metropolitan Reporting Areas (MRA). For those reports, initially, we performed preliminary descriptive analysis that showed diverse mean count for different MRAs and syndrome. In improve accuracy, we stratified the MRA into three levels, based on the mean daily count: 1–19, 20–49 and $\geqslant 50$.

2.2. Control-chart-based algorithms

BioSense uses a rate-based version of the C2 algorithm, one of the three algorithms (C1, C2 and C3) developed for the Early Aberration Reporting System (EARS) [10,11]. Our previous study [5] also showed consistent improved sensitivity to simulated signals when the length of the sliding baseline was increased from 7 days to 14 or 28 days. For comparison to regression models, we applied the count-based and rate-based C2 algorithm using 14-day and 28-day baselines, with the descriptive names C2c14, C2r14, C2c28 and C2r28.

Let L be length of baseline and let the index day be the day whose data are being tested. Analogous to the sliding baseline used to compute regression model coefficients, the index day expected value for syndromic visits was calculated using the 14-or 28-day baseline, i.e. $L = 14$ or 28, separated from the index day by a 2-day buffer. For the count-based C2 methods, the expected value is simply the mean baseline count. For the rate method, the expected value is:

$$Expected_{indexday} = N_{indexday} \times \frac{\sum_{i=1}^{L} n_i}{\sum_{i=1}^{L} N_i} \quad (1)$$

where n_i is number of visits meeting a syndrome definition on baseline day i, N_i is number of total visits, $N_{indexday}$ is number of total visits on the test day.

For normalization of mean deviations, a dispersion measure for rate-based C2 was calculated by

$$SD_{indexday} = \frac{\sum_{i=1}^{L} \left| n_i - N_i \times \frac{\sum_{j=1}^{L} n_j}{\sum_{j=1}^{L} N_j} \right|}{L} \quad (2)$$

2.3. Regression-based models

The regression modeling was guided by the following findings from [7]. First, seasonality is captured using indicator variables rather than from models based on multiple years of data history that are unavailable for many data streams. These indicator variables are given a fixed value for each day of a time interval intended to represent current seasonal behavior. Several interval lengths were tested, and a 14-day interval gave the best forecasts and is used in the models below. Second, sliding baselines for the regression inference were limited to 56 days; longer baseline intervals did not improve forecast accuracy. Third, use of total visits for surrogate rates gave substantial forecast improvements in respiratory and gastrointestinal syndrome data for some time series, and the models in the Methods use the total visits in several ways. Fourth, the day-of-week indicator variables improved forecasts for some syndromes and were used in the models tested. Experience with these ED data series shows weekly patterns for certain syndrome groups, but holiday and post-holiday effects are not consistent and were not modeled.

Our five regression models controlled for day of week (dow) with indicator variables and seasonality with indicator variables for 14-day time period. Three of the models also controlled for total daily visits, which includes both baseline and the indexday (test

Download English Version:

https://daneshyari.com/en/article/517737

Download Persian Version:

https://daneshyari.com/article/517737

Daneshyari.com