Journal of Biomedical Informatics 57 (2015) 446–455

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Practical comparison of aberration detection algorithms for biosurveillance systems

Hong Zhou [a,*], Howard Burkom [b], Carla A. Winston [c], Achintya Dey [a], Umed Ajani [a]

[a] Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333, United States
[b] Johns Hopkins Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723, United States
[c] Veterans Health Administration, Office of Public Health Surveillance and Research, 3801 Miranda Avenue (132), Palo Alto, CA 94304, United States

ARTICLE INFO

Article history:
Received 11 December 2014
Revised 25 August 2015
Accepted 26 August 2015
Available online 31 August 2015

Keywords:
Aberration detection
Algorithm
Control chart
Poisson regression
Biosurveillance

ABSTRACT

National syndromic surveillance systems require optimal anomaly detection methods. For method performance comparison, we injected multi-day signals stochastically drawn from lognormal distributions into time series of aggregated daily visit counts from the U.S. Centers for Disease Control and Prevention's BioSense syndromic surveillance system. The time series corresponded to three different syndrome groups: rash, upper respiratory infection, and gastrointestinal illness. We included a sample of facilities with data reported every day and with median daily syndromic counts $\geqslant 1$ over the entire study period. We compared anomaly detection methods of five control chart adaptations, a linear regression model and a Poisson regression model. We assessed sensitivity and timeliness of these methods for detection of multi-day signals. At a daily background alert rate of 1% and 2%, the sensitivities and timeliness ranged from 24 to 77% and 3.3 to 6.1 days, respectively. The overall sensitivity and timeliness increased substantially after stratification by weekday versus weekend and holiday. Adjusting the baseline syndromic count by the total number of facility visits gave consistently improved sensitivity and timeliness without stratification, but it provided better performance when combined with stratification. The daily syndrome/total-visit proportion method did not improve the performance. In general, alerting based on linear regression outperformed control chart based methods. A Poisson regression model obtained the best sensitivity in the series with high-count data.

Published by Elsevier Inc.

1. Introduction

Syndromic surveillance systems have been widely introduced since the year 2000 to enhance public health situational awareness and for detecting and tracking disease outbreaks [1]. One of the major challenges for these systems is to identify events of interest from substantial "background noise" in surveillance data http://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm. Automated surveillance systems use statistical aberration detection methods to identify increases above predetermined thresholds in monitored medical encounters or other healthcare-seeking data classified into broad clinical categories, denoted as *syndromes*. Typically, systems form time series by aggregating healthcare visit counts for each day for each syndrome of interest [2]. Analysts then apply statistical methods to test these series prospectively for anomalies that might be indicators of health concerns. It is critical to select the optimal aberrancy-detection algorithms to detect disease outbreaks and public health threats.

For more than a decade, the statistical methods C1–C3 of the Early Aberration Reporting System (EARS) of the Centers for Disease Control and Prevention (CDC) have been among the most widely used globally because of their simplicity and ease of use [3]. In their original form, these methods were designed strictly for count data. They employed sliding baselines for some seasonality adjustment but made no other adjustment for systematic data behavior. Later modifications extended the applicability of the C2 method [4]. Although sliding baselines for some seasonality adjustment have been employed, both known sources of biases (e.g., weekly patterns and holidays) and unknown or difficult-to-catch biases (e.g., reporting delay/error and weather) may still exist. Various adjustments to reduce bias have been suggested in the aberration detection practice. C2 can be used on proportion, i.e., daily syndrome count/daily total visits. However, previous studies have indicated no advantage of proportion method on aberration detection [4]. Thus, total visits baseline adjustment (e.g., 28 days) has been used for enhancing the EARS C2 algorithms [4]. In addition,

* Corresponding author. Tel.: +1 404 498(6293); fax: +1 404 498(6870).
E-mail address: fwd6@cdc.gov (H. Zhou).

http://dx.doi.org/10.1016/j.jbi.2015.08.023
1532-0464/Published by Elsevier Inc.

weekday/weekend stratification has been applied in C2 methods to adjust for weekday/weekend effect [4]. In regression modeling, total visits and (or) day-of-week and (or) seasonal terms have also been used as covariates or offset in models [5,6]. The advantage of stratification adding upon the total visit adjustment is unclear, however.

A critical limitation for assessing different statistical methods in detecting and tracking disease outbreaks is the lack of detailed information about target signals, i.e. of time series labeled according to the effects of known outbreaks. Most studies used simulated background data [5] which may not represent the real situations of time series for various syndromes. In addition, many studies [4,6] looked at detection performance using single day injected signals that may not represent real epidemic curves. Ideally, authentic background data from daily surveillance should be used with realistic multiday signals in comparison of aberration detection method.

Hospital/clinic facilities are the frontline invaluable sources for outbreak detection since outbreaks usually start at local level [7]. Previous studies mainly have focused on populations at the large metropolitan area, county, or greater [6,8]. Localized clusters of interest (e.g., within a facility or a group of facilities) could have insufficient size to be detected in analyses when data were available only at a large region, such as a county or even state level. In addition, lack of specific localized information might make the targeted investigation, prevention, and intervention difficult. CDC's national biosurveillance system, originally known as BioSense, has evolved significantly since first becoming operational in 2005. Data from the U.S. Department of Veterans Affairs' Veterans Health Administration (VHA) provide a unique source of actual time series at the facility level to the CDC BioSense surveillance system. We compared five control chart based methods, a linear regression model and a Poisson regression model on the performance of aberration detection. Objectives of this study are to answer the following practical questions: (1) which methods yield the best detection performance, in terms of sensitivity and alerting timeliness? (2) How effective are the techniques of weekday/weekend stratification and total-visit adjustment? (3) Can the total-visit adjustment eliminate the need to stratify, or should these techniques be combined? (4) Do advantages persist for smaller facilities whose aggregated records have diminished time series structure? So far, simultaneous comparisons of detection performance of various control chart methods and regression modelings with different stratification and adjustment strategies applying to the same time series are lacking. It is important to answer the above four questions in more practical situations through detecting aberrations by simulating realistic signals in authentic data streams of multiple scales of activities.

2. Methods

We used daily syndrome counts of several syndromic time series in outpatient VHA facilities as baseline data and added multiday data effects of simulating events of disease outbreaks.

2.1. Baseline data

Currently, hospitals, outpatient clinics, and public health departments across the United States provide data to BioSense [http://www.cdc.gov/nssp/biosense/index.html]. The number of participating jurisdictions and facilities has varied. In 2008, BioSense received daily data streams from up to 532 civilian hospitals, 333 hospitals and clinics from the U.S. Department of Defense, and more than 770 facilities of the U.S. Department of Veterans Affairs' Veterans Health Administration (VHA) [9]. All these facilities are monitoring an evolving collection of syndromes and subsyndromes. The number of involved facilities has increased over time.

We used daily syndrome counts of outpatient visits at the VHA healthcare facility level as baseline data. The daily counts were derived by classifying patient records into syndrome groups according to diagnosis codes based on the International Classification of Diseases, 9th Revision [9]. We included records from facilities that reported every day from January 2010 through May 2011 with median daily syndromic counts $\geqslant 1$ over the entire study period. To examine the algorithm detection performances over a variety of data scales and seasonal behaviors, we selected three BioSense syndromes: rash, upper respiratory infection (URI), and gastrointestinal (GI) [4]. The rash and GI syndromes were chosen for typical small and large counts of records from most facilities, while the URI syndrome counts represented counts with a strong seasonal pattern. Many VHA facilities had few outpatient visits on holidays as well as weekends. Therefore, the 14 federal holidays during the report period were recoded as Sundays for purposes of stratified algorithm calculations and for testing the day-of-week effect in the regression models.

We used data from a 56-day baseline period for the comparison methods. The baseline period for the first test day began on 1/1/2010; hence, the test period was from 2/28/2010 through 5/31/2011. We used a "sliding" baseline to reflect the recent data behavior, so that each baseline period ended two days before the date of concern. The purpose of the two-day buffer was to avoid contamination of the baseline data with a potential early phase of an outbreak [6].

With these restrictions of consistent and non-sparse reporting, the data for this study were from 62 facilities in 39 states. We calculated median daily count during the study period for each facility syndrome. Based on experience and on published literature [6,10] on alerting algorithm performance for algorithm assessment and comparison, we categorized facilities into three median daily count categories (1–4, 5–9, and at least 10) for each syndrome. We assessed five control chart-based algorithms (C2 Count, C2 Proportion, CuSUM Count, C2 Count Adjusted, and CuSUM Count Adjusted) and two regression models (Linear Reg and Poisson Reg) (Supplementary Material). The methods were applied to these separate categories to compare the sensitivity and timeliness of aberration detection by these methods at different levels of sparseness in authentic data.

We stratified the baseline days used into weekdays and weekend days to calculate μ (E_t) and SD (SD'). The 56-weekday/weekend-stratified baseline days contained $\sim$40 weekdays and $\sim$16 weekends. Each of the five control chart-based algorithms was tested with and without this stratification. Tokars et al. [4] enhanced the performance of EARS C2 algorithms by lengthening the baseline periods from seven days to 14 and 28 days. We chose 56 baseline days to ensure an accurate and stable calculation of μ and SD after the stratification.

The two regression models were run separately for each facility and syndrome, with the expected value for each index day predicted from the regression model applied to the 56 days of data preceding the index day with a two-day buffer. The standard deviation (SD) of the expected value was calculated by using the equation

$$SD = \frac{\sum_{i=1}^{56} |n_i - E_i|}{56}$$

where n_i is the observed syndrome count and E_i is the model-expected syndromic count for baseline day i. The regression test statistic was computed by using the equation

$$\text{Reg} = \frac{(X_t - E_t)}{SD}$$