Evaluation of a Syndromic Surveillance System Using the WSARE Algorithm for Early Detection of an Unusual, Localized Summer Outbreak of Influenza B: Implications for Bioterrorism Surveillance*

Zalman Kaufman MSc[1], Weng-Keen Wong PhD[6], Tamar Peled-Leviatan MSc[1], Erica Cohen MD MPH[4], Chana Lavy BA[4], Gali Aharonowitz MD MSc[1], Rita Dichtiar BA[1], Michal Bromberg MD MPH[1], Ofra Havkin MD MPH[5], Ehud Kokia MD MHA[2,3] and Manfred S. Green MD MPH PhD[1,3]

[1]Israel Center for Disease Control, Ministry of Health, Tel Hashomer, Israel
[2]Maccabi Health Care Services, Tel Aviv, Israel
[3]Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv, Israel
[4]Rehovot Subdistrict Health Office, Ministry of Health, Rehovot, Israel
[5]Central District Health Office, Ministry of Health, Ramleh, Israel
[6]Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Key words: syndromic surveillance, evaluation, outbreak, WSARE, detection algorithm

Abstract

Background: Syndromic surveillance systems have been developed for early detection of bioterrorist attacks, but few validation studies exist for these systems and their efficacy has been questioned.

Objectives: To assess the capabilities of a syndromic surveillance system based on community clinics in conjunction with the WSARE algorithm in identifying early signals of a localized unusual influenza outbreak.

Methods: This retrospective study used data on a documented influenza B outbreak in an elementary school in central Israel. The WSARE algorithm for anomalous pattern detection was applied to individual records of daily patient visits to clinics of one of the four health management organizations in the country.

Results: Two successive significant anomalies were detected in the HMO's data set that could signal the influenza outbreak. If data were available for analysis in real time, the first anomaly could be detected on day 3 of the outbreak, 1 day after the school principal reported the outbreak to the public health authorities.

Conclusions: Early detection is difficult in this type of fast-developing institutionalized outbreak. However, the information derived from WSARE could help define the outbreak in terms of time, place and the population at risk.

IMAJ 2007;9:3–7

Several syndromic surveillance systems have been developed in recent years for early detection of bioterrorist attacks [1-9]. However, few validation studies have been published for these systems and questions have been raised about their efficacy [1,10-13]. Since actual bioterrorist incidents have not occurred, validating these systems and the algorithms for early detection embedded within them is problematic [7]. Two approaches are used to evaluate the performance of syndromic surveillance systems [7,11]. One approach is to simulate a variety of different pathogen-oriented bioterrorist events where the simulated data are superimposed on a series of real data streams. The alerting algorithms use these data streams as a test case for evaluation. In the second approach, the ability of the alerting algorithms is tested retrospectively on documented disease outbreaks, based on the data routinely assembled in the surveillance system.

During the last few years the Israel Center for Disease Control has been engaged in the development of a syndromic surveillance system as part of a national bioterrorism preparedness program. This system is based partly on data from community clinics. Among the several analytic tools that we considered for integration into our system was the WSARE (What's Strange About Recent Events) data mining algorithm. For retrospective evaluation purposes, we used an unusual outbreak of influenza B that occurred in June 2004 in an elementary school in central Israel.

Materials and Methods

A localized outbreak, characterized by flu-like illness, was identified at the beginning of June 2004 in one of two elementary schools in a small town in the center of the country, with fewer than 10,000 inhabitants (referred to as town X). The children of two adjacent fourth grade classes were most affected by the outbreak. At noon on June 2, the school principal reported the outbreak to the regional health department, based on unusual absenteeism in those two classes. That evening the parents of 35 of the 59 pupils in the two classes were interviewed by public health officials. Two days previously (May 31), 28 pupils were reported to be sick. The epidemic curve based on these partial data is shown in Figure 1. The three most prevalent symptoms reported by the parents were headache (100%),

* This paper was presented at the 2004 Syndromic Surveillance Conference, 3–4 November 2004, Boston, Massachusetts.

WSARE = What's Strange About Recent Events
HMO = health management organization

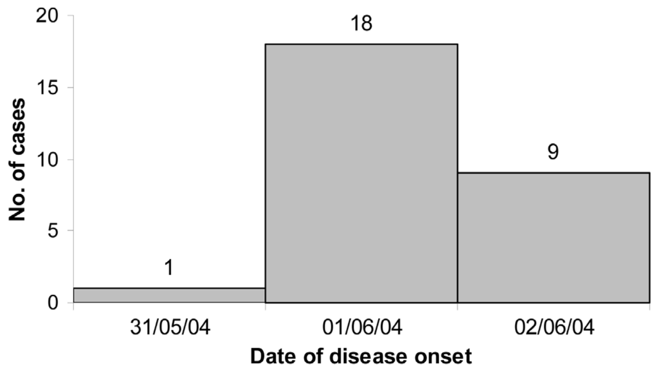

Figure 1. A partial epidemic curve for pupils in two 4th grade classes of an elementary school in town X (n=28).

fever (75%) and sore throat (46.4%). On June 6, pupils of that school were requested to come to the local clinics for a throat swab. On June 8 a laboratory diagnosis of influenza type B was confirmed.

Data available at the ICDC

We restricted the retrospective evaluation to community clinics as the sole data source. The four health management organizations in Israel have clinics around the country. In accordance with the National Health Insurance law, every individual in the country is registered in one of the four HMOs. Maccabi Health Care Services, the second largest HMO and serving about 25% of the population, shares its data with the Israel Center for Disease Control surveillance system. Maccabi has developed and implemented a computerized information system that is fully employed at all levels of the organization. Demographic and clinical data are collected in real time from all levels of care and stored in a central database. Data from individual records of patient visits to Maccabi's network of clinics throughout the country are transferred daily to the ICDC. Each record contains the following fields: patient code (a code that cannot breach personal privacy), date of birth, date of clinic visit, symptoms/diagnosis coded according to the ICD-9-CM, city code, census tract code if available, and zip code. Only records with ICD-9 codes pertaining to infectious diseases – like those of the respiratory system, the gastrointestinal system and other less specific, such as fever and viral infections – are transferred. The databases at the ICDC are managed by the SAS package software, version 8.2. At present the ICDC surveillance system is not functioning on a daily basis, so its performance in this study is evaluated retrospectively.

Analytic tool

The WSARE 3.0 algorithm [14] was implemented in this evaluation to detect anomalies relevant to the outbreak under study. WSARE is an anomalous pattern detection system, developed by the Auton Lab's team at Carnegie Mellon University, specifically for use with medical data [15]. Briefly, WSARE seeks answers to the question "What's strange about recent events?" WSARE approaches this problem using a rule-based anomaly detection approach. This program takes as input a date-indexed biosurveillance data stream such as emergency department data. It then operates over a range of dates, for each of which it compares events from that day with events from a baseline distribution obtained from past data [See Appendix for more details on WSARE].

Data preparation

Individual records of patient visits from 40 cities and towns under the jurisdiction of the regional health department where the outbreak occurred were used for the evaluation. The starting date was 2 May and the data set ended on 11 June. In total, the data set included 17,821 records. The analysis was limited to four fields included in the individual records: date of visit to the clinic, city code, ICD-9 code, and age as a categorical variable (the age variable was grouped into nine categories: < 2, 2–5, 6–14, 15–24, 25–34, 35–44, 45–54, 55–64, and 65 years and older). An extra field was added to the data set, where visit dates were converted to corresponding days of the week and then grouped into three categories: Sundays, Fridays, other days. The WSARE algorithm takes into consideration inherent patterns that could be the result of seasonality or days of the week. We included this extra field since a clear weekly pattern can be portrayed in the community healthcare data, where maximum visits occurred on Sundays and minimum visits on Fridays. The first 10,000 records, corresponding to the first 3 weeks of the above period (2 to 23 May), were defined as records needed to create the baseline by the WSARE algorithm.

Results

When the WSARE algorithm was applied to the regional data set of patient visits to Maccabi clinics, six statistically significant anomalies were identified for the period 24 May to 11 June 2004 [Table 1]. Of those six, two consecutive highly significant anomalies were detected in town X, where the outbreak occurred. Apart from the city code, these two anomalies, the first detected on June 2 and the second on June 3, comprised two other equivalent components: the 6–14 age category and the ICD-9 code of viral

Table 1. Statistically significant anomalies detected by the WSARE algorithm in Maccabi's data set from May 24 until June 11

Date	Today's ratio	Other ratio	Town	Age	Diagnosis	*P*
25 May 2004	2.55	0.14	Z	–	URT	< 0.0001
27 May 2004	2.35	0.16	Z	–	URT	< 0.0001
2 June 2004	1.44	0.02	X	6–14	VI	< 0.0001
3 June 2004	2.78	0.02	X	6–14	VI	< 0.0001
6 June 2004	4.08	1.83	–	–	CHP	0.002
7 June 2004	2.45	0.29	Y	–	VI	0.002

URT = upper respiratory tract infections, VI = viral infections, CHP = chickenpox

ICDC = Israel Center for Disease Control
HMO = health management organization
ICD-9-CM = International Classification of Diseases, 9th revision, clinical modification

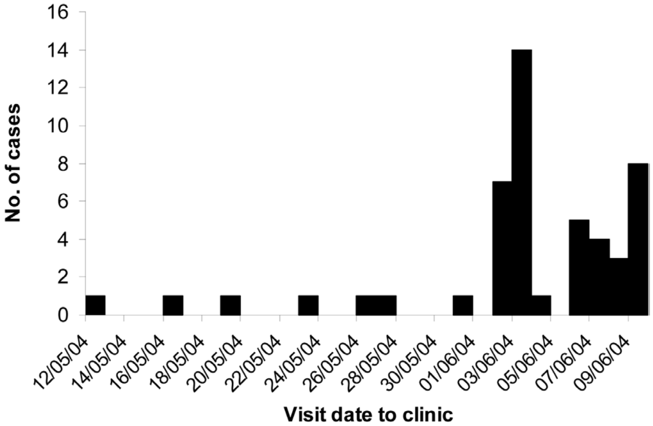

Figure 2. Visits of children aged 6–14 to the Maccabi clinic where the school is situated, and diagnosed as suffering from viral infections. 5 June 2004 is Saturday; the community clinics were closed on that day.

infection. The 6–14 age category includes the age of 10 years old, which is the characteristic age of grade 4 pupils, and viral infection was the most prevalent diagnosis given by physicians at the local Maccabi clinic to pupils of that school at the start of the outbreak, as documented during an investigation by the regional health department.

When the WSARE algorithm was applied again, focusing only on patients residing in town X and using age as a continuous variable, the same two anomalies were detected, this time pointing exactly to children aged 10 years. Based on the two additional attributes of the anomalies identified in town X, Figure 2 was constructed, describing visits to Maccabi clinics by children aged 6–14 diagnosed as suffering from viral infection from May 12 until June 9. Seven and 14 cases were documented on June 2 and June 3, respectively, compared with none to one visit during the preceding 3 weeks. There is a self-explanatory shift of 1 day between the pattern of visits to a Maccabi clinic observed in Figure 2 and that of the epidemic curve [Figure 1] constructed based on data of dates of disease onset gathered for about 60% of the pupils in the two fourth grade classes.

Of the other four anomalies identified, one anomaly of June 7 also had the ICD-9 code of viral infection as one of its attributes. This anomaly was identified in town Y, which is in close proximity to town X. This anomaly, identified later on the time axis in comparison to those anomalies of town X, could result from the outbreak in town X spreading to adjacent towns and villages.

Discussion

Through implementation of the WSARE algorithm on a historical regional data set of one of the four HMOs operating in Israel, two successive significant anomalies on June 2 and June 3 were detected in town X, comprising two additional attributes that characterized the outbreak in its early stages. The fact that two consecutive anomalies were detected, and not just one, increases the confidence that the signal is true. This type of reasoning was also used by officials running the syndromic surveillance system in New York City [16]. They concluded that anomalies warranting further consideration were those that were sustained for at least 2 days or were detected in two different data sources.

Currently, data from Maccabi are available for analysis 24 hours after a patient visit is documented by a physician. Taking into account this delay, the first anomaly could be detected only on June 3, if the monitoring is operative on a regular basis. This was one day after the school principal had actually notified the regional health department about the outbreak. If data were available for analysis close to real time, this gap – between the time the outbreak was already evident on the ground and the time needed for detection through syndromic surveillance – could be shortened. We speculate that in specific scenarios such as the studied outbreak, the practical benefits from using syndromic surveillance for early detection are limited.

School absenteeism data and other non-traditional data are being monitored by several syndromic surveillance systems [7,11,17]. The inclusion of non-traditional data sources within this kind of system is based on the assumption that monitoring behavior trends will enable earlier outbreak detection. In one study, using a simulated outbreak of tularemia, school absenteeism data were shown to be a major contributor to the early detection of the abnormality [7]. We tried to assess the theoretical contribution of school absenteeism data in detecting the studied outbreak. As found during the regional health department investigation, the absenteeism in the two grade 4 classes in which the outbreak started was double by June 1, before any visits of the schoolchildren to at least two of the three clinics operating in the town were documented. Theoretically, incorporating these data within the surveillance system could lead to early detection of the event. However, computerization of school absenteeism data is not regular practice in Israel, and if the data were computerized they usually would not be available in real time. A time lag of at least 24 hours until these kinds of data are available for analysis within the surveillance system seemed quite realistic. Based on this assumption, integration of school absenteeism data within the surveillance system, at least on this occasion, would not result in earlier detection. However, these data could strengthen the confidence in the anomalies identified in the medical data sets.

The source of the outbreak and its unusual characteristics have not yet been resolved. Although highly unlikely, this type of outbreak could, theoretically, be the result of an intentional act. Bioterrorist incidents can take several forms, with institutionalized outbreaks being one of the possible scenarios. Based on the data presented so far, the role of syndromic surveillance in early detection of institutionalized fast-developing outbreaks seems quite limited. It was quite easy for the school officials to detect the outbreak, already in its second day, due to the unusual absenteeism noted at the school. In addition, the detection capabilities of syndromic surveillance systems during winter could deteriorate. In fact, we superimposed the outbreak data on winter background data sets and detection was much more problematic (data not shown). Reingold and others [1,10-13] have voiced quite a skeptical attitude towards the early-detection capabilities of syndromic surveillance. Nevertheless, Reingold thought it quite plausible that a preexisting syndromic surveillance system can help define in a timely manner the affected or at-risk population in need of preventive measures. We agree with this insight. A more comprehensive syndromic surveillance system, where data are available for analysis in real time, could reassure the public health officials dealing with the influenza outbreak in town X, already on the evening of June 2 or during the morning of June 3, that the investigated event was localized and not widespread (not even to the other elementary school in the town). For example, later on June 7, the regional public health department could have been notified that an anomaly was detected at a nearby town (town Y), suggesting that the outbreak is spreading. In retrospect, no reports of unusual morbidity were received at the regional health department from town Y at that time. This anomaly could have been a signal of a small unrelated event.

Conclusions

Although early detection is hard to achieve in this type of outbreak, the prompt and reliable information produced by a centralized syndromic surveillance is of great value for supporting outbreak management. This would apply to both bioterrorist incidents and natural outbreaks. Outbreaks such as described in this paper can serve as testbeds for evaluating syndromic surveillance systems, while giving analysts and public health officials the opportunity to experience with signals produced by this kind of system.

References

1. Bravata DM, McDonald KM, Smith WM, et al. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. *Intern Med* 2004;140:910–22.
2. Lewis MD, Pavlin JA, Mansfield JL, et al. Disease outbreak detection system using syndromic data in the Greater Washington DC area. *Am J Prev Med* 2002;23:180–6.
3. Lazarus R, Kleinman K, Dashevsky I, et al. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerg Infect Dis* 2002;8:753–60.
4. Lober WB, Karras BT, Wagner MM, et al. Roundtable on bioterrorism detection: information system-based surveillance. *J Am Med Inform Assoc* 2002;9:105–15.
5. Lober WB, Trigg LJ, Karras BT, et al. Syndromic surveillance using automated collection of computerized discharge diagnoses. *J Urban Health* 2003;80(Suppl 1):i97–106.
6. Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response early aberration reporting system (EARS). *J Urban Health* 2003;80(Suppl 1):i89–96.
7. Lombardo J, Burkom H, Elbert E, et al. A system overview of the Electronic Surveillance System for the Notification of Community-Based Epidemics (ESSENCE II). *J Urban Health* 2003;80(Suppl 1): i32–42.
8. Irvin CB, Nouhan PP, Rice K. Syndromic analysis of computerized Emergency Department patients' chief complaints: an opportunity for bioterrorism and influenza surveillance. *Ann Emerg Med* 2003;41:447–52.
9. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. *Emerg Infect Dis* 2004;10:858–64.
10. Reingold A. If syndromic surveillance is the answer, what is the question? *Biosecur Bioterror* 2003;1:1–5.
11. Stoto MA, Schonlau M, Mariano LT. Syndromic surveillance: is it worth the effort? *Chance* 2004;17:19–24.
12. Bravata DM, Sundaram V, McDonald KM, et al. Evaluating detection and diagnostic decision support systems for bioterrorism response. *Emerg Infect Dis* 2004;10:100–8.
13. Sosin DM. Syndromic surveillance: the case for skillful investment. *Biosecur Bioterror* 2003;1:247–53.
14. Wong WK, Moore A, Cooper G, Wagner M. Bayesian anomaly pattern detection for disease outbreaks. In: Proceedings of the International Conference on Machine Learning; 2003:808–15.
15. The Auton project. URL: http://www.autonlab.org/autonweb/showSoftware/159/
16. Mostashari F. Syndromic surveillance in New York City. Presentation at the CDC International Conference on Emerging Infectious Diseases, Atlanta, 2002. [Cited January 2003] Available from: URL: ftp://ftp.cdc.gov/pub/infectious_diseases/iceid/2002/pdf/mostashari.pdf
17. Tsui F-C, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM. Technical description of RODS: a real-time public health surveillance system. *J Am Med Inform Assoc* 2003;10:399–408.

Correspondence: Z. Kaufman, Israel Center for Disease Control, Gertner Institute for Health Policy Research, Sheba Medical Center, Tel Hashomer 52621, Israel.
Phone: (972-3) 737-1500
Fax: (972-3) 534-9881
email: zalman.k@icdc.health.gov.il

Appendix A – The WSARE algorithm

The What's Strange About Recent Events (WSARE) algorithm detects anomalous patterns in discrete, multidimensional data sets with a temporal component. WSARE requires records within a specified temporal period to be defined as *recent* records. Records preceding the recent data points in time are used to produce a *baseline* data set that represents normal behavior. WSARE compares the recent data against the baseline data to find the most significant change in recent records. This change is described using a rule that is made up of components of the form $X_i = V^j_i$, where X_i is the *i*th feature and V^j_i is the *j*th value of that feature. For example, the one-component rule *Gender* = *Male* characterizes the subset of the data involving males. These rules should not be interpreted as rules from a logic-based system in which the rules have an antecedent and a consequent. Rather, these rules can be thought of as SQL SELECT queries. Like SQL SELECT queries, rules in WSARE can consist of multiple components connected by a logical AND. For example, a two-component rule could be *Gender* = *Male* AND *Home Location* =

NW, which identifies the group consisting of males living in the northwest region of the city. Each rule identifies the proportion of records in the data that match the rule. WSARE finds the group with the most significant change in its proportion between the recent data set and the baseline data set.

The WSARE algorithm consists of three steps. First of all, the baseline data set is created. Secondly, the algorithm searches for the best scoring rule using both the recent and the baseline data sets. Finally, a *P* value for the best scoring rule is calculated using a randomization test. We will briefly provide more details about the three steps above. For a complete description of the WSARE algorithm we refer readers to two articles [1,2]. Furthermore, there are a number of variations on the WSARE algorithm that differ only in how the baseline data set is created. We will be describing WSARE 3.0, which uses a Bayesian network to create the baseline data set.

Creating the baseline distribution is a difficult task due to temporal trends in the data such as those caused by seasonal and day-of-week variations. WSARE 3.0 accounts for these trends by using a Bayesian network to model the joint probability distribution of the features of the data. These features can be divided into *environmental attributes*, which are features such as the season and the day of week that cause trends in the data, and *response attributes*, which are the remaining features such as age and gender. WSARE 3.0 learns a Bayesian network from all data that precede the recent period. During this Bayesian network structure learning phase, environmental attributes are prevented from having parents because we are not interested in predicting their distributions, but, rather, we want to use them to predict the distributions of the response attributes. Once the Bayesian network structure is learned, we can then produce a conditional probability distribution that represents the baseline behavior given the environmental attributes for the current day. As an example, suppose we are monitoring Emergency Department data and that the environmental attributes Season, Day of Week, and Weather cause fluctuations in these data. Also, let the response attributes be $X_1, \ldots, X_n$. Assuming that today is a snowy winter Saturday, we can use the joint probability distribution captured by the Bayesian network to produce the conditional probability distribution $P(X_1,\ldots,X_n \mid Season = Winter, Day\ of\ Week = Saturday, Weather = Snow)$, which intuitively represents the baseline distribution given the conditions for the current day. The baseline data set can consequently be produced by sampling a large number of records from this conditional probability distribution.

Once the baseline data set is generated, we need to search for the best scoring rule, which characterizes the group with the most unusual shift in proportions between the baseline and recent data sets. We start by examining all one-component rules, which consist of a single feature-value combination such as *Home Location* = NW. In order to score each rule, we need to obtain the number of records that match the rule and the number that do not match the rule in the recent data set. We also need to obtain similar counts for the baseline data set. With these four counts, we set up a two-by-two contingency table as shown in Table 1. We then run Fisher's exact test on the table below to obtain a *P* value. This *P* value will be referred to as the *score* in order to distinguish it from the *P* value calculated in the next step. Note that more unusual rules have lower scores. At this point, the best one-component rule has been found. We will refer to the best one-component rule as BR^1. The algorithm then attempts to find the best two-component rule for the day by adding on one extra component to BR^1 through a greedy search. This extra component is determined by supplementing BR^1 with all possible feature-value pairs, except for the one already present in BR^1, and selecting the resulting two-component rule with the best score. Scoring is performed in the exact same manner as before, except that the counts are calculated from the records that match the two-component rule. In order to prevent overfitting, additional components are only added to the best rule so far if the addition of those components is statistically significant. Also, we usually limit the maximum number of components on a rule to be two due to the computational expense of the search.

Table 1. A sample two-by-two contingency table as used by WSARE to score rules

	C_{recent}*	$C_{baseline}$
Home Location = NW	6	496
Home Location ≠ NW	40	9504

* The heading C_{recent} indicates a count from the recent data set while $C_{baseline}$ indicates a count from the baseline data set.

Let BR be the best scoring rule found and let *Score*(BR) be the score of BR. We cannot interpret *Score*(BR) as its actual *P* value because the process for finding the best scoring rule involves multiple hypothesis tests. The final step of WSARE accounts for the multiple hypothesis testing problem by calculating a compensated *P* value for the best scoring rule through a randomization test in which the date and the remaining features are assumed to be independent. The randomization test consists of several iterations, typically around 1000. On iteration *j*, we shuffle the dates between records in the recent and the baseline data sets to produce a randomized data set called DB^j_{rand}. We then find the best scoring rule BR^j on DB^j_{rand}. At the end, we determine where *Score*(BR) would be ranked among the values of *Score*(BR^j) from all the iterations. The compensated *P* value CPV is calculated as:

$$CPV = \frac{\#\text{ of times } Score(BR^j) < Score(BR)}{\#\text{ randomization test iterations}}$$

Finally, an alarm is sounded if the compensated *P* value CPV is lower than a threshold, which is typically set to the standard threshold value of 0.05.

References

1. Wong WK, Moore A, Cooper G, Wagner M. Bayesian network anomaly pattern detection for disease outbreaks. In: *Proc Int Conf Machine Learn* 2003;808–15.
2. Wong WK. Data Mining for Early Disease Outbreak Detection. Doctoral Dissertation, School of Computer Science, Carnegie Mellon University, 2004.s