Comparing Aberration Detection Methods with Simulated Data

Lori Hutwagner,* Timothy Browne,*
G. Matthew Seeman,* and Aaron T. Fleischauer*

We compared aberration detection methods requiring historical data to those that require little background by using simulated data. Methods that require less historical data are as sensitive and specific as those that require 3–5 years of data. These simulations can determine which method produces appropriate sensitivity and specificity.

The Early Aberration Reporting System (EARS) was developed to allow analysis of public health surveillance data. Several alternative aberration detection methods are available to state and local health departments for syndromic surveillance. Before 2001, most statistical aberration detection methods required at least 5 years of background data (1–6). However, with the release of *Bacillus anthracis* in the U.S. mail shortly after the September 11, 2001, World Trade Center attacks, substantial interest has emerged in public health tools that could be rapidly implemented without requiring years of background data. Newly developed nonhistorical aberration detection methods can require as little as 1 week of data to begin analysis, although they have not been extensively evaluated against traditional historical methods (7,8).

The objective of our study was to determine the sensitivity, specificity, and time to detection of 3 methods that require <3 years of historical baseline data, C1–MILD (C1), C2–MEDIUM (C2), and C3–ULTRA (C3), and compare the results with those of 2 methods that require 5 years of historical baseline, the historical limits method (2) and the seasonally adjusted cumulative sum (CUSUM) (5), based on simulated data. Simulated data were used to avoid some of the interpretation difficulties that can come from making these comparisons on the basis of empirically observed, natural disease data. All 5 of these methods are components of EARS (7).

The Study

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. All 3 methods are based on a positive 1-sided CUSUM calculation. For C1 and C2, the CUSUM threshold reduces to the mean plus 3 standard deviations (SD). The mean and SD for the C1 calculation are based on information from the past 7 days. The mean and SD for the C2 and C3 calculations are based on information from 7 days, ignoring the 2 most recent days. These methods take into consideration daily variation because the mean and SD used by the methods are based on a week's information. These methods also take seasonality into consideration because the mean and SD are calculated in the same season as the data value in question.

Since 1989, results from the historical limits method have been used to produce Figure 1 in the Morbidity and Mortality Weekly Report. This method compares the number of reported cases in the 4 most recent time periods for a given health outcome with historical incidence data on the same outcome from the preceding 5 years; the method is based on comparing the ratio of current reports with the historical mean and SD. The historical mean and SD are derived from 15 totals of 3 intervals (including the same 4 periods, the preceding 4 periods, and the subsequent 4 periods over the preceding 5 years of historical data).

The seasonally adjusted CUSUM method is based on the positive 1-sided CUSUM where the count of interest is compared to the 5-year mean and the 5-year SD for that period. The seasonally adjusted CUSUM was originally applied to laboratory-based *Salmonella* serotype data.

To calculate sensitivity, specificity, and time to detection, all 5 detection methods of EARS were used to independently analyze 56,000 sets of artificially generated case-count data based on 56 sets of parameters. These 56 sets of parameters each generated 1,000 iterations of 6 years of daily data, 1994–1999, by using a negative binomial distribution with superimposed outbreaks. Means and standard deviations were based on observed values from national and local public health systems and syndromic surveillance systems. Examples of the data included national and state pneumonia and influenza data and hospital influenzalike illness. Adjustments were made for days of the week, holidays, postholiday periods, seasonality, and trend. Any 6 years could be used, but the years 1994–1999 were used to set day of the week and holiday patterns and to avoid any problems that programs might have with the year 2000. Fifty (89%) of these datasets then had outbreaks superimposed throughout the data. Three types of outbreaks were used, each representing various types of naturally occurring events: log normal, a rapidly increasing outbreak; inverted log normal, a slowly starting outbreak; and a single-day spike. These types of outbreaks were combined with different SDs and incubation times to create 10 different types of outbreaks that had equal probability of being included in the simulated data. A year of final simulated data can be seen in the Figure, with original data and

*Centers for Disease Control and Prevention, Atlanta, Georgia, USA

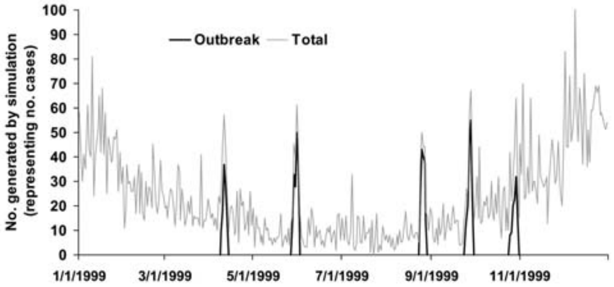

Figure. Example of 1 year of simulated data with simulated outbreaks. Simulated data are based on real means and standard deviations with different types of simulated outbreaks randomly inserted.

outbreaks that were added. As a result of these analyses, the statistically marked aberrations, or flags, produced by the 5 detection methods were evaluated for their specificity, sensitivity, and time to detection. These data can be obtained at http://www.bt.cdc.gov/surveillance/ ears/datasets.asp.

In our study, sensitivity was defined as the number of outbreaks in which ≥1 day was flagged, divided by the total number of outbreaks in the data. An outbreak was defined as a period of consecutive days in which varying numbers of aberrant cases were added to the baseline number of cases. An outbreak had days before and after it when no aberrant cases were added to the baseline case counts. Specificity was defined as the total number of days that did not contain aberrant cases (and that were not flagged), divided by the total number of days that did not contain aberrant cases. Based on these definitions, actual values for sensitivity and specificity were calculated.

Time to detection was defined as the number of complete days that occurred between the beginning of an outbreak and the first day the outbreak was flagged. For example, if a method flags an outbreak on the first day, its time to detection is 0. Likewise, if it flags on the second day, its time to detection is 1, and so on. Time to detection is an average of the times to detection for each outbreak and dataset. Only outbreaks that were flagged on at least 1 day were included in the average. Therefore, sensitivity is needed to completely interpret time to detection. We calculated 2-sided 95% confidence values, and they were relatively small and consistent.

Overall, the CUSUM methods (the seasonally adjusted CUSUM, C1, C2, and C3) had similar times to detection, but their sensitivities varied (Table). Specifically, C1, C2, and C3 showed increasing sensitivity from 60% to 71% to 82%, respectively. The seasonally adjusted CUSUM and C3 methods had similar sensitivities, 82.5% and 82.3%, but C3 had a higher specificity, 88.7% and 95.4%. The historical limits and C1 and C2 methods showed varying sensitivities (44%–71%), with C1 and C2 having the highest, but all demonstrated similar specificities (96%–97%).

When results were stratified by outbreak type, 1-day outbreaks (i.e., spikes) exhibited the lowest sensitivities. Analysis was broken down by dataset and outbreak type (online Appendix Tables 1 and 2, available at http://www.cdc.gov/ncidod/EID/vol11no02/04-0587_app1.htm and http://www.cdc.gov/ncidod/EID/vol11no02/04-0587_app2.htm).

For the 6 datasets that contained noise but no outbreaks, no sensitivity or time to detection exist to calculate. The overall specificity for the seasonally adjusted CUSUM, historical limits, C1, C2, and C3 were 88.7%, 98.3%, 97.2%, 97.2%, and 95.2%, respectively. The specificity for these 6 datasets was consistent with general results. The historical limits method showed superior specificity in all but the last dataset.

Conclusions

These simulations demonstrate that the methods for aberration detection that require little baseline data, C1, C2, and C3, are as sensitive and specific as the historical limits and seasonally adjusted CUSUM methods. As expected, C1, C2, and C3 showed increasing sensitivities in accordance with their intended sensitivity levels (C1 being the least sensitive, C3 being the most), but with decreasing specificities as sensitivity increases. Seasonally

Table. By method, overall sensitivity and specificity and time to detection

Type of method	Name	Sensitivity (%)	Specificity (%)	Time to detection (d)*
Historical methods (at least 5 y historical data)	Seasonally adjusted CUSUM†	82.5	88.7	1.272
	Historical limits‡	43.9	96.3	2.942
Nonhistorical methods (<3 y historical data)	C1–MILD§	60.1	97.0	1.122
	C2–MEDIUM¶	71.2	97.0	1.319
	C3–ULTRA**	82.3	95.4	1.307

*Time to detection must be interpreted with sensitivity because time to detection does not include missed outbreaks.

†The seasonally adjusted CUSUM method sums the positive differences of the current value from the mean for a period similar to the current value over 5 years.

‡The historical limits method compares the current sum of 4 time periods to the mean of the sum of 15 totals of 4 time periods surrounding the current point of interest over 5 years.

§The C1–MILD method is based on CUSUM, but the calculations reduce to the current value being greater than the mean plus 3 standard deviations (SD), with the mean and SD based on the past 7 days.

¶The C2–MEDIUM method is based on CUSUM, but the calculations reduce to the current value being greater than the mean plus 3 SD, with the mean and SD based on the past 7 days shifted by 2 days.

**The C3–ULTRA method is based on CUSUM, summing the positive difference of the current value from the mean for 3 days, with the mean and SD based on the past 7 days shifted by 2 days.

adjusted CUSUM and the historical limits method also showed sensitivities and specificities as expected, with the seasonally adjusted CUSUM having the lower specificity and higher sensitivity. These findings emphasize the effectiveness of aberration detection methods without requiring long-term historical data as a baseline.

Since the 10 simulated outbreaks were randomly generated by using consistent rates, the sensitivity, specificity, and time to detection could be stratified by dataset and outbreak type. The results of these analyses were largely congruent with the expected findings, with some variations. The simulated datasets are designed for public health officials to select a dataset that best reflects their data of interest or the type of outbreak they are anticipating to determine which method provides them with the sensitivity and specificity they would find useful. The simulated datasets can also be used to make comparisons with other methods.

The aberration detection methods C1, C2, and C3 are used in several states, counties, and local public health departments. Public health departments are able to apply these methods to data sources that do not have long periods of baseline data. Public health departments are also able to apply 1 set of methods they understand to various types of diseases, covering different frequencies and seasonalities. The C1, C2, and C3 methods have detected outbreaks of public health interest, including West Nile disease and the start of the influenza season.

C1, C2, and C3 demonstrate consistency over the various situations represented in these simulations. Other aberration detection methods exist, as do other simulated datasets. The simulated datasets presented in this paper cover a larger variety of types of data that might be expected in public health. These simulated datasets also include enough past years of data so that methods that require 5 years of historical information can also be used in the comparisons. These simulations provide a method to fairly compare other methods among themselves and to the methods included in EARS.

The simulations were based on means and SDs to help determine which method performs better under which circumstances. When deciding which method to use, the potential user should base the decision on the sensitivity or specificity or the time to detection.

A potential limitation is that the method for calculating average times to detection disregards undetected outbreaks. Therefore, times to detection should not be considered without also taking into account the sensitivity. However, this method was preferred over the alternative of assigning arbitrary numbers of days to detection for outbreaks that were not detected since the alternative method could lead to misinterpretation of the data. Another limitation is that the artificial datasets may not fully reproduce the nuances of natural disease occurrences. While approximations, the simulated data were generated based on naturally observed data and included variations for trend over time, days of the week, seasons, and holidays. Therefore, while these comparisons represent relative sensitivities, specificities, and times to detection, we do not know whether results using naturally occurring data would be consistent.

The results of this study suggest that the EARS historical methods do not have a strong advantage when compared with nonhistorical methods. In fact, the lack of historical data does not impair the EARS outbreak detection methods. This study also demonstrates the effectiveness of artificial outbreak data in comparing and evaluating outbreak detection methods. As aberration detection methods are increasingly being used by state and local health departments to monitor for naturally occurring outbreaks and bioterror events, this study contributes to the quest to determine the most efficient method for analyzing surveillance data.

Ms. Hutwagner works with the Bioterrorism Preparedness and Response Program at the Centers for Disease Control and Prevention on developing aberration detection methods for their national “drop-in surveillance” system and ongoing syndromic surveillance. She has been implementing these methods at various sites in the United States and internationally.

References

1. Teutsch SM, Churchill RE, editors. Principles and practice of public health surveillance. New York: Oxford University Press; 2000.
2. Stroup DF, Williamson GD, Herndon JL, Karon J. Detection of aberrations in the occurrence of notifiable diseases surveillance data. Stat Med. 1989;8:323–9.
3. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A statistical algorithm for the early detection of outbreaks of infectious disease. J R Stat Soc Ser A Stat Soc. 1996;159:547–63.
4. Simonsen L, Clarke JM, Stroup DF, Williamson GD, Arden NH, Cox NJ. A method for timely assessment of influenza-associated mortality in the United States. Epidemiology. 1997;8:390–5.
5. Hutwagner LC, Maloney EK, Bean NH, Slutsker L, Martin SM. Using laboratory-based surveillance data for prevention: an algorithm for detecting salmonella outbreaks. Emerg Infect Dis. 1997;3:395–400.
6. Stern L, Lightfoot D. Automated outbreak detection: a quantitative retrospective analysis. Epidemiol Infect. 1999;122:103–10.
7. Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). J Urban Health. 2003;80:i89–96.
8. Hutwagner L, Thompson W, Groseclose S, Williamson GD. An evaluation of alternative methods for detecting aberrations in public health surveillance data. American Statistical Association, Joint Statistical Meetings, Proceedings of the Biometrics Section. Indianapolis; 2000 Aug. p. 82–5.

Address for correspondence: Lori Hutwagner, Centers for Disease Control and Prevention, 1600 Clifton Rd, Mailstop C18, Atlanta, GA 30333, USA; fax: 404-639-0382; email: lhutwagner@cdc.gov