Comparing Historical Limits Method with Regression Model for Weekly Monitoring of Notifiable Diseases

Hong Zhou[1], Howard Burkom[2], Susan Katz[1], Ruth Jajosky[1], Willie Anderson[1], Achintya Dey[1], Umed Ajani[1]
[1]Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333
[2]Johns Hopkins Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723

Abstract

Every week, the Centers for Disease Control and Prevention (CDC) monitors counts of more than 75 notifiable diseases. In order to detect aberrations of numbers of occurrences, detection methods must be robust across different diseases and allow for endemic variation. To compare methods, we injected stochastic lognormal-distributed signals into each of 12 selected diseases' weekly time series of newly reported case counts from the CDC's National Notifiable Diseases Surveillance System. We used provisional data (before end-of-year reconciliation with state health departments) from 2006–2010 as baseline and from 2011–2014 for testing. We compared the Historical Limits Method (HLM) to a method derived from quasi-Poisson regression model (England method), using both 1- and 4-week baseline data units for testing each method. Both methods allowed for seasonal effects by calculating empirical thresholds using corresponding weeks in past years' data. At a 2% background alert rate, mean sensitivity for signal detection ranged from 25–78% for short signals (peaking at 1–2 weeks) and from and 50–88% for long signals (peaking at 3–5 weeks). With 1-week data units, sensitivities to detect short signals were higher and alerting delays were lower than with 4-week data units for both methods. The England method outperformed HLM regardless the length of signals and weeks of data units.

Key Words: aberration detection, Poisson regression, historic limit method, disease surveillance

1. Introduction

The ability to detect aberrant clusters of reportable infectious disease quickly and accurately for meaningful response is a central goal of public health institutions [1-3]. Application of automated statistical techniques to detect possible outbreaks is particularly important in national disease surveillance systems that serve large populations and receive a high volume of reports because manual review and investigation of all reports are not always feasible. The National Notifiable Diseases Surveillance system (NNDSS), operated by the Centers for Disease Control and Prevention (CDC), in collaboration with the Council of State and Territorial Epidemiologists (CSTE), collects incidence data on more than 75 nationally notifiable diseases from the 50 U.S. states, New York City, Washington D.C., and five U.S. territories on a weekly basis. It provides an important source of infectious disease surveillance data for the United States.

Since 1980s, the historical limits method (HLM) [4] has been used in NNDSS to detect unusually high or low numbers of reported cases and to indicate changes in long-term

trends of reported cases [5]. Farrington et al. [3] developed a quasi-Poisson regression model (England method) for monitoring weekly data in the 1990s for the Communicable Disease Surveillance Centre, which is part of the National Public Health Service for Wales. This method is widely applied in European countries in disease surveillance systems [6]. To account for seasonal effects by calculating empirical thresholds using data from the corresponding weeks of past years, HLM compares the number of reported cases in the current *4-week period* for a given health event with historical data from the preceding 5 years. However, the England method uses *weekly* data in the regression. A previous study [1] indicates that HLM lacks adjustment for long-term trend, year-to-year variation, and outliers (disease clusters and aberration). Although a few studies [2, 7, 8] have examined whether adjustment with regression models provides better alerting by controlling for predicted behaviors, none of these studies examined the effects of length of data analysis unit and the duration of aberrations on the detection performance of HLM and regression models.

In this study, we performed a systematic comparison between the HLM and the England method. We used a set of weekly case counts of provisional (before end-of-year reconciliation with state health departments) disease reports from CDC NNDSS as baseline data and added realistic simulated data effects of disease aberrations. We selected diseases that varied by expected volume of cases, seasonality, endemic behavior, historical trend and other characteristics. We compared the detection performance of HLM with the England method by challenging them with both sudden and slow-growing aberrations, testing with both the 1- and 4-week units of analysis. We compared background alert rate, sensitivity, and alerting delay under various endemic conditions and made recommendations based on our findings.

2. Methods

2.1 Baseline data

State and selected local health departments send case notifications of nationally notifiable diseases to CDC weekly throughout the year. These provisional data are published in MMWR weekly. After the year ends, staff in state health departments finalize reports of cases for the year with local or county health departments and reconcile the data with reports that were previously sent to CDC throughout the year. The finalized data are published in the MMWR Summary of Notifiable Diseases, United States [5]. Since the purpose of our study was to compare methods of rapid detection of aberrations, we used the provisional weekly data in this report. NNDSS data from 2006 through 2014 were used with the first 5-year period (all of 2006–2010 inclusive) as the initial baseline and 1/1/2011 through 12/31/2014 as test period. We selected 12 diseases as examples: Chickenpox (Varicella), Coccidioidomycosis, Cryptosporidiosis, Giardiasis, Hepatitis A, Legionellosis, Listeriosis, Lyme disease, Meningococcal disease, Pertussis, Salmonellosis, and Shigellosis [9]. Chickenpox (Varicella), Salmonellosis, and Lyme disease were chosen as typical of high weekly records counts. The Hepatitis A, Meningococcal disease, and Listeriosis were used to represent typically low counts. Chickenpox (Varicella) and Hepatitis A counts had a downward long-term trend while Pertussis counts had an upward trend. Time series for these selected diseases had various seasonal patterns, except for Coccidioidomycosis, Hepatitis A, and Listeriosis.

2.2 Historical Limits Method

In HLM, the predicted reported count for the current 4-week period is the mean of the reported number of cases during the preceding 15 4-week periods, the corresponding 4-

week periods, and the following 4-week periods, for the previous 5 years [4]. The use of comparable 4-week periods from the past 5 years is intended to account for seasonality. In the below diagram, X_0 is the current 4-week period, X_1, X_2 and X_3 are the preceding, corresponding and following 4-week periods for last year, respectively. Thus, the mean of the 15 weekly counts from X_1 through X_{15} is used as the expected count for the current period, and their standard deviation is a measure of expected spread.

2011		X_0	← "Current" Four Weeks	
2010	X_1	X_2	X_3	
2009	X_4	X_5	X_6	
2008	X_7	X_8	X_9	
2007	X_{10}	X_{11}	X_{12}	
2006	X_{13}	X_{14}	X_{15}	

2.3 England Method

The England method is a quasi-Poisson distribution model that assumes that variance is a function of the mean. It allowed for seasonal variation by calculating empirical thresholds for the current week using comparable weeks in the past 5 years [3,6].

2011				W_0	← "Current" Week		
2010	W_1	W_2	W_3	W_4	W_5	W_6	W_7
2009	W_8	W_9	W_{10}	W_{11}	W_{12}	W_{13}	W_{14}
2008	W_{15}	W_{16}	W_{17}	W_{18}	W_{19}	W_{20}	W_{21}
2007	W_{22}	W_{23}	W_{24}	W_{25}	W_{26}	W_{27}	W_{28}
2006	W_{29}	W_{30}	W_{31}	W_{32}	W_{33}	W_{34}	W_{35}

In above diagram, W_0 is the current week of year 2011, W_1, W_2, W_3 , W_4, W_5, W_6, and W_7 are the proceeding 3 weeks, current week and following 3 weeks for last year, respectively. W_1 through W_{35} compose the 35 baseline values from the past 5 years. The England method includes a time term that is measured in weeks to account for long-term trend: $E_t = \beta_0 + \beta_1 * time$

2.4 Threshold Calculation

Usually, HLM uses 4-week data units and the England method uses 1-week data units. However, to ensure thorough testing, we ran the two methods for both 1- and 4-week data units. For both methods, we took the following approach for selecting baselines, so that the baselines would be comparable: For 1-week units, we used 35 baseline weeks (7 consecutive weeks in each of the past 5 years), as in the England method. For 4-week units, we used the 15 baseline 4-week periods in past 5 years, as in the HLM.

For each disease, an alarm is triggered if the count for the current 1- or 4-week period is larger than a calculated threshold. The threshold calculation is $threshold = E_{current} + 2 * SD_{current}$, where $E_{current}$ is the predicted value of the current 1- or 4-week period, and $SD_{current}$ is the standard deviation of the predicted value calculated by using the equation

$$SD_{current} = \frac{\sum_{i=1}^{j} |n_i - Ei|}{j}$$

where j=15 and 35 for 4- and 1-week data units, respectively; n_i is the observed disease count, E_i is the predicted value for each baseline 1- or 4-week period *i.(* i.e., mean of previous observed count in HLM and predicted value by quasi-Poisson regression in England method).

2.5 Outbreak Signal Simulation

We used simulated signals to compare the detection performance of the two alerting methods. We generated these signals from lognormal random draws because the incubation periods of many infectious diseases are lognormally distributed [10, 11].

For each disease, our signal simulation process was to form short series of weekly outbreak-attributable counts for addition to the authentic data. To obtain detectable signals to challenge the alerting methods, we set the estimated number of cases in the peak outbreak week to $M = 2 * SD_{background}$, where $SD_{background}$ is the standard deviation of the disease-specific background counts. For a two-parameter lognormal distribution, we used two sets of location and shape parameters $\zeta = 1.0$, $\delta = 0.15$ (for short signals) and $\zeta = 1.8$, $\delta = 0.3$ (for long signals) calculated experimentally using literature on infectious disease incubation [12]. From the lognormal probability density, we used the peak-week count M to calculate the total number N of attributable cases, generated N incubation periods for these cases using lognormal draws, and rounded these periods to the nearest week. Then we summed to find the number of injected cases for each week after the start of the signal. For each disease, this procedure was used to generate injected signals beginning at each chosen target week.

2.6 Method Evaluation

The inclusion of past clusters or aberration in historical data may introduce bias in the method's ability to detect aberrations. To reduce this bias, we trimmed extreme outliers by truncating values greater than 4 standard deviations above the baseline mean to the mean itself [1]. Thus, we created evaluation datasets by adding simulated signals onto the trimmed time series. For each disease, we added with two series of 80 consecutive injected signals, one set of short signals and one set of long signals. For the first signal, we chose a theoretical onset date of Jan 1, 2011. The starting weeks of each subsequent signal varied according to exact lengths of previous signals, which ranged from 2 to 5 weeks for short signals and 6 to 15 weeks for long signals. We repeatedly used the initial time series for each disease to ensure that 80 injected signals were tested for both short and long signals. We applied the two aberration detection methods to the evaluation datasets for each disease. We calculated the background alert rate, sensitivity, and alerting delay for detecting injected signals using these three steps: (1) running each method to estimate predicted value, standard deviation, and threshold (=$2*SD_{current}$ above predicted value) without signal injection for each disease week; (2) identifying weeks during which the observed count exceeded the threshold values and calculating background alert rate; and (3) identifying weeks during which the value of observed count plus injected counts exceeded the threshold values and calculating sensitivity and alerting delay.

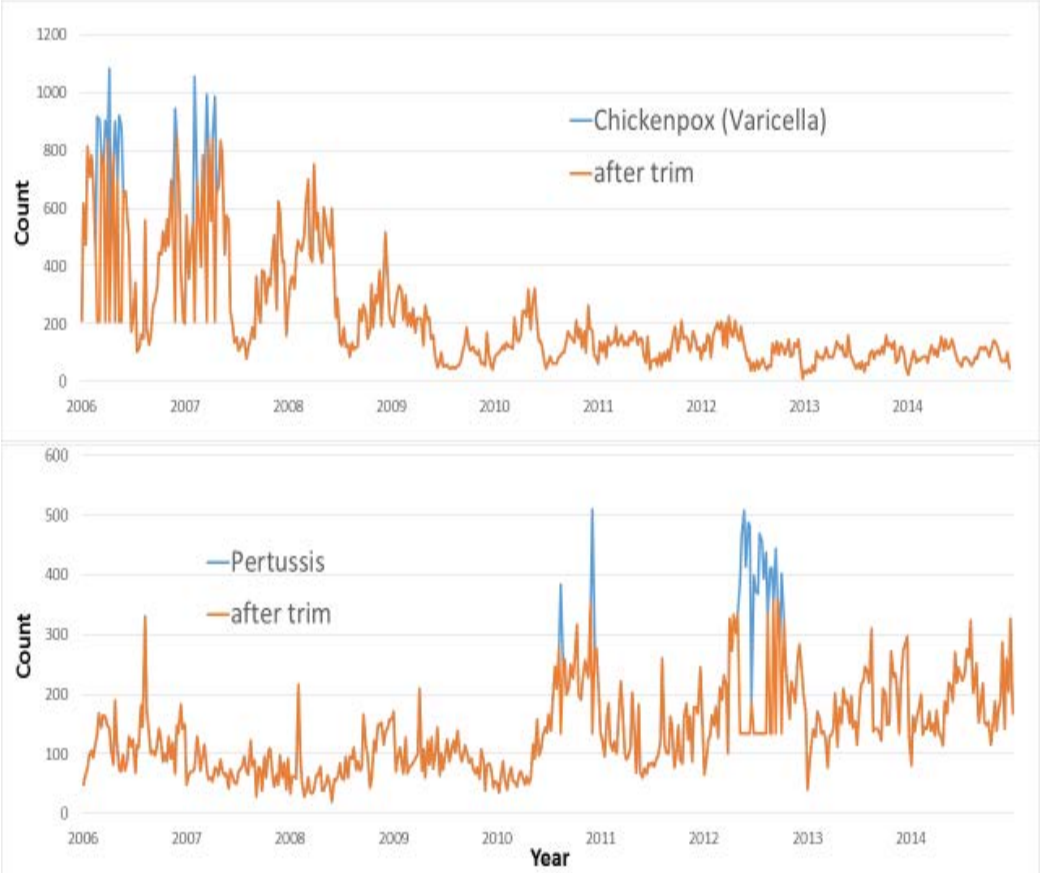
In our study, the background alert rate was calculated as the ratio of weeks when baseline count exceeded the threshold to the total number of tested weeks. We refer to a *background alert rate* rather than a *false alert rate* (=1-specificity) because there was no way to identify and exclude real outbreaks from the data [13]. The calculation of specificity would require the unverifiable assumption that true outbreaks are absent from the baseline data. The sensitivity was defined as the proportion of the number of signals detected before the lognormal distribution peak to the total number of injected signals, with the rationale that detection is required no later than the outbreak peak. Alerting delay was calculated as the number of weeks from the start of injection to the first algorithm alert occurring not later than the peak inject week. If there was no alert or an alert occurred after the peak week, the alerting delay was set as 1 plus the mean of peak week of signals injected in the disease counts. For example, if the mean of peak week of injected signals for a disease is 3 and the peak week of a specific signal is during week 2, and a method alerts this signal on the second week, its alerting delay is 2 weeks. But, if it alerts after the second week or does not alert at all, its alerting delay is 4 (1+3) weeks. We calculated the means of background alert rate, sensitivity, and alerting delay and their minimum and maximum values to indicate the variation across the 12 diseases. We used ANOVA to compare the two aberration detection methods in backgrounds alert rates, sensitivities, and timeliness adjusting for data unit and signal length. SAS 9.3 was used to perform all analyses.

3. Results

3.1 Descriptive Data

The baseline data we used were weekly disease counts from NNDSS provisional data from 2006–2014. The 12 selected diseases represented a variety of volumes of disease counts, trends, and seasonal behaviors. For example, the weekly counts for time series of Chickenpox (varicella) had a declining long-term trend and a strong seasonality (Figure 1). Varicella also showed year-to-year variation and some outliers (blue lines in the figure). The time series of Pertussis had an upward trend. Pertussis also showed strong year-to-year variation and some outliers, but not as consistent a seasonal pattern as varicella. The orange lines in Figure 1 are the baseline data after trimming. We used the baseline data after trimming in the method comparison.

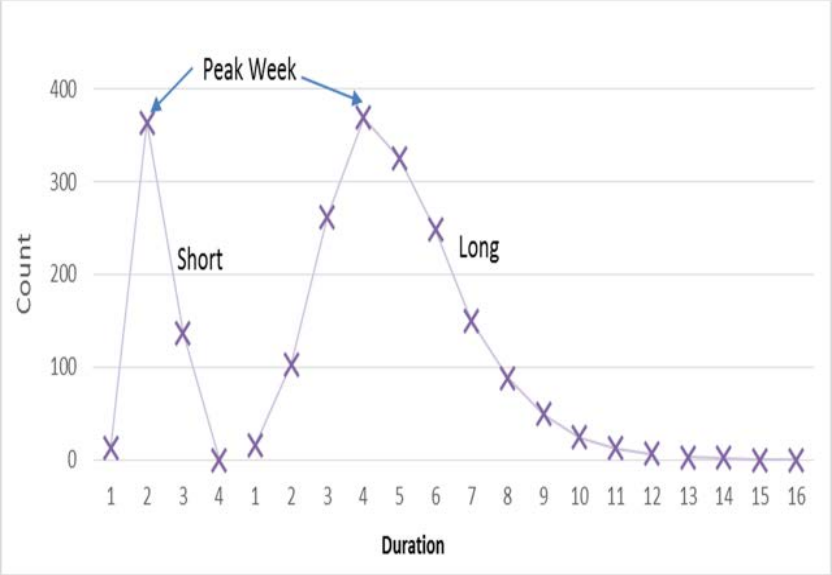
Figure 1. Examples of weekly counts for time series of selected diseases from NNDSS provisional data 2006-2014



3.2 Characteristics of Injected Signals

As described above, we set peak-week counts for injected signals at two standard deviations above the baseline mean to obtain practical detection challenges. We tested 80 short signals and 80 long signals for each disease from Jan 1, 2011 through Dec. 31, 2014. The durations of the long signals ranged from 6 to 15 weeks with peak weeks ranging from 3 to 5 weeks. The durations of the short signals ranged from 2 to 5 weeks with peak weeks ranging from 1 to 2 weeks (Figure 2).

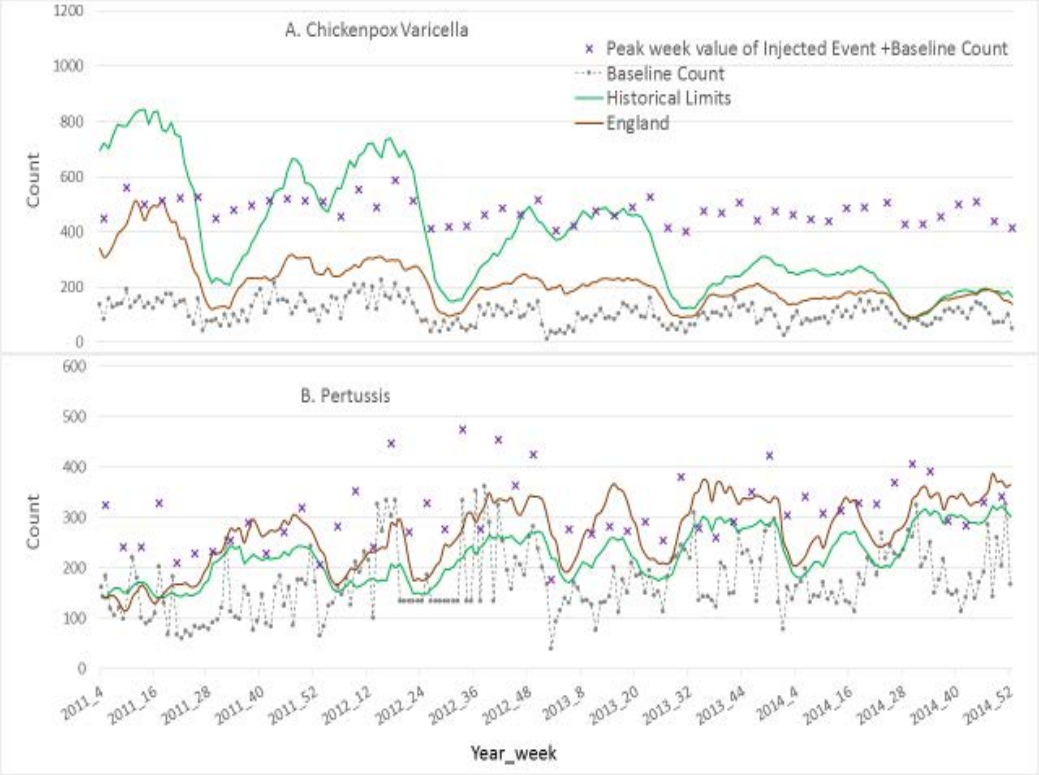
Figure2. Examples of Injected Signals



3.3 Threshold Comparison

Our method of evaluation was based on the threshold of two standard deviations (SD) above the predicted value generated by each method. An example of thresholds and the created evaluation datasets with series of consecutive injected signals from 2011 through 2014 is displayed in Figure 3 for Chickenpox (varicella) and Pertussis. To avoid crowding the plot with symbols, only the peak week value of each signal is shown. For Chickenpox (Panel A), the threshold is high for HLM, especially in the first three years. In contrast, the threshold from the England method is relatively stable from year to year. On the other hand, if the disease (e.g., Pertussis) has an upward trend (Panel B), the HLM tends to generate lower threshold than the England method.

Figure 3. Threshold comparison of methods for example diseases with 1-week data unit (2011-2014)



3.4 Background Alert Rate, Signal Sensitivity, and Alerting Delay

The mean background alert rates, sensitivities, alerting delay for the 12 diseases and their minimal (Min) and maximal (Max) at two SD above predicted value are presented in Table 1. They are stratified according to signal length (no signal for background alert rate) and data units.

For both methods, using 1-week data units resulted in a lower (better) background alert rate than using 4-week data units. The HLM model has slightly lower background alert rate than the England method, but has more variation across the 12 diseases than the England method based on differences between maximum and minimum values.

The sensitivity is higher (better) when using 1-week data units than 4-week units, and the advantage of using 1-week data units is even larger for short signals. Both methods have better sensitivity for detection of longer signals than short signals. The England method has higher sensitivity than does HLM regardless of length of signals or 1- or 4-week data units.

Using 1-week data units yields a shorter alerting delay than 4-week data units for all signals and methods. The differences in average delay time are 0.7 week (~5 days) for long signals and 0.4 week (~3 days) for short signals. The England method has shorter alerting delay than HLM regardless of length of signals or 1- or 4-week data units, but the differences are smaller than the differences between 1-week vs. 4-week data units.

Table1. Background alert rate, sensitivity and alert delay of detection injected signal with 12 diseases baseline series from NNDSS national provisional data (2011 - 2014) at threshold of 2*standard deviation above predicted value

Outcome	Signal (Peak week)	Data Unit	Historic Limits			England		
			Mean	Min	Max	**Mean**	Min	Max
Background alert rate (%)	N/A	1-week	**6**	0	21	**6**	0	12
		4-week	**8**	0	27	**10**	0	20
Sensitivity (%)	Short (1-2)	1-week	**67**	21	93	**78**	48	99
		4-week	**25**	0	54	**34**	16	58
	Long (3-5)	1-week	**79**	53	98	**88**	61	99
		4-week	**50**	4	84	**66**	35	88
Alerting delay (Week)	Short (1-2)	1-week	**2.0**	1.3	2.5	**1.9**	1.3	2.5
		4-week	**2.4**	1.9	2.9	**2.3**	1.7	2.8
	Long (3-5)	1-week	**3.1**	2.3	3.9	**2.8**	2.2	3.9
		4-week	**3.8**	2.9	4.5	**3.5**	2.7	4.4

ANOVA tests show that the estimated background alert rate was similar between HLM and England methods ($p=0.641$). However, adjusting for data unit and signal length, England method had significantly higher sensitivity ($p=0.003$) and shorter alerting delay ($p=0.029$) than HLM method. As expected, 1-week data unit yield significantly higher sensitivity ($p<0.001$) and shorter alerting delay ($p<0.001$) than 4-week data unit.

4. Discussion

The purpose of this study was to compare HLM with the England method for aberration detection in national disease surveillance systems. In our study, using weekly provisional counts of 12 selected diseases, we found that both methods have better performance in detecting long signals than short signals. The use of 1-week data units yields consistently better sensitivity and shorter detection delays than conventional 4-week data units. Using 1-week data units has even greater advantage for detecting short signals. The England method gives better sensitivity and alerting delay than HLM for the same duration of signals and data units.

HLM lacks adjustment for long-term trend and yearly variation. The simple calculation of the predicted value as the mean of the 15 baseline data points does not account for the long-term trend and adjusts poorly for year-to-year variation and outliers. The England method's quasi-Poisson distribution is more appropriate for most disease count data. In addition, its adjustment for long-term trends is important for many diseases. The quasi-Poisson (England) model yields better adjustment for year-to-year variation and thus provides more representative thresholds.

From our study, using the 1-week data unit gave higher sensitivity and shorter detection delay than using 4-week data units. The advantage of 1-week data units was even greater for detecting short signals. A plausible explanation is that injected signals were diluted by

the 4-week data units. The 1-week units also yielded a shorter background alert rate, possibly because the 4-week data units decreased the standard deviation and then generated relatively lower thresholds. The choice of 4-week data units in original HLM was to control the weekly fluctuation in disease reporting that is usually due to irregular reporting rather than to disease incidence [14]. However, in public health practice, some outbreaks may last only a few weeks, and thus the use of 4-week data units may miss or delay the detection of these events.

NNDSS surveillance data provide valuable information on trends in infectious disease incidence for the United States. These data are also used to detect sudden changes in disease occurrence and distribution. Provisional weekly data are often reported as early as possible to alert public health practitioners regarding emerging problems. However, it should be noted that increases in the number of reported cases of a particular disease could also be a result of batched reporting related to a jurisdiction's priorities and practices, changes in physician reporting due to increased awareness, or changes in case finding due to screening or modified diagnostic methods. Completeness of reporting may vary among jurisdictions and may relate to the condition or disease being reported. Although we used weekly provisional data in this study to perform method evaluation on the best available timely data, future studies could assess predictive performance from provisional counts relative to final, corrected counts.

There are several additional limitations of this study. First, we tested the methods only at the national level. Therefore, our results might not apply to data at state, county, or city levels having much smaller disease counts. Second, the 12 selected diseases used for the study do not represent all disease characteristics. Data derived for other diseases may differ in scale, seasonality, and other systematic behaviors from the study series. Third, the simulated signals should not be considered representative of all authentic signal types. Though data effects of true outbreaks are rarely available in quantities required for statistical significance, such data should be used for testing whenever possible.

The results of our study indicate that compared to traditional HLM, the England method performs better for aberration detection in NNDSS. If HLM is to be used, 1-week data units instead of current 4-week data units would likely improve detection performance.

References

1. *Levin-Rector, A., et al., Refining historical limits method to improve disease cluster detection, New York City, New York, USA. Emerg Infect Dis, 2015.* ***21****(2): p. 265-72.*
2. *Choi, B.Y., et al., Comparison of various statistical methods for detecting disease outbreaks. Computational Statistics, 2010.* ***25****(4): p. 603-617.*
3. *Farrington, C.P., et al., A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. Journal of the Royal Statistical Society. Series A (Statistics in Society), 1996.* ***159****(3): p. 547-563.*
4. *Stroup, D.F., et al., Detection of aberrations in the occurrence of notifiable diseases surveillance data. Statistics in Medicine, 1989.* ***8****(3): p. 323-329.*
5. *Notifiable Diseases and Mortality Tables. Morbidity and Mortality Weekly Report (MMWR), 2015.* ***64(34);ND-598-ND-615****.*
6. *Hulth, A., et al., Practical usage of computer-supported outbreak detection in five European countries. Euro Surveill, 2010.* ***15****(36).*

7. *Jackson, M.L., et al., A simulation study comparing aberration detection algorithms for syndromic surveillance. BMC Med Inform Decis Mak, 2007.* ***7****: p. 6.*
8. *Hong Z, Burkom H, Winston C, Dey A, Ajani U. Practical comparison of aberration detection algorithms for biosurveillance systems.* Journal of Biomedical Informatics*. 2015;57:446-455.*
9. *Adams, D.A., et al., Summary of notifiable diseases--United States, 2012. MMWR Morb Mortal Wkly Rep, 2014.* ***61****(53): p. 1-121.*
10. *Philippe, P., Sartwell's incubation period model revisited in the light of dynamic modeling. J Clin Epidemiol, 1994.* ***47****(4): p. 419-33.*
11. *Sartwell, P.E., The distribution of incubation periods of infectious disease. 1949. Am J Epidemiol, 1995.* ***141****(5): p. 386-94; discussion 385.*
12. *Detrick, F., Medical Management of Biological Casualties. U.S. Army Medical Research Institute of Infectious Diseases, Sept. 2000.*
13. *Tokars, J.I., et al., Enhancing time-series detection algorithms for automated biosurveillance. Emerg Infect Dis, 2009.* ***15****(4): p. 533-9.*
14. *Stroup, D.F., et al., Evaluation of a Method for Detecting Aberrations in Public Health Surveillance Data. American Journal of Epidemiology, 1993.* ***137****(3): p. 373-380.*