A Comparison of Regression Models with Adaptive Control-chart Algorithms for Aberration Detection in Biosurveillance Systems

Hong Zhou[1], Howard Burkom[2], Carla Winston[3], 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

[3]Veterans Health Administration, Office of Public Health Surveillance and Research, 3801 Miranda Avenue (132), Palo Alto, CA 94304

Abstract

Biosurveillance systems require robust anomaly detection methods. For detection-method performance comparisons, we injected multi-day lognormal distributed signals into gastrointestinal (GI) syndrome-related time series of aggregated daily counts from the Centers for Disease Control and Prevention's BioSense syndromic surveillance system. CDC is part of the U.S. Department of Health and Human Services. We included a sample of facilities with data reported every day and with median daily syndromic counts ≥ 3 over the entire study period from Jan. 2010 through May 2011. We compared alerting algorithms on the basis of a Poisson regression model (including covariates for day of week, total visits, and seasonality) with two adaptations of the cumulative sum (CuSUM) chart and three adaptations of the Shewhart chart (with variations on whether/how to adjust for total visits). We assessed sensitivity and timeliness of these methods for detection of injected multi-day signal events. Sensitivity was defined as the ratio of number of events detected before the event peak to the total number of injected signals. Alerting timeliness was calculated as the number of days from the start of injection to the first algorithm alert not later than the peak day of the injected signal. At a daily background alert rate of 1%, the sensitivities and timeliness measured as delay (in days) before signal detection ranged from 26%-66% and 2.6-3.3 days, respectively. We also examined sensitivity and timeliness with and without stratification by weekday versus weekend/ holiday. For time series with mean daily counts ≥ 10, the Poisson regression-based method achieved higher sensitivity and slightly shorter mean detection delays than the chart-based methods for detection of multi-day signals in GI-related visit series.

Key Words: aberration detection, algorithm, Poisson regression, biosurveillance

1. Introduction

The Centers for Disease Control and Prevention (CDC) established the BioSense program with the intent of providing real-time biosurveillance for early aberration detection [1]. The Early Aberration Reporting System (EARS) is used by CDC and worldwide by numerous organizations' surveillance systems for finding statistical anomalies that may signal disease outbreaks. Currently, hundreds of hospitals and public health departments across the United States provide data to BioSense for routine monitoring and anomalies alerting that might signal disease outbreaks.

Due to their simplicity, EARS C2 methods are widely used at CDC and worldwide [6]. However, these basic methods do not adjust for common systematic data behaviors, and the resulting biases make them suboptimal for many syndromic time series [2]. Sources of biases that are both known (e.g., weekly patterns and holidays) and unknown or difficult-to-model (e. g., reporting delay/error and weather effects) complicate time series monitoring. Thus, adjustment for total visits at baseline and weekday/weekend stratification have been used to enhance performance of the EARS C2 algorithms [2].

Total visit counts and (or) day-of-week have been used for adjustment in regression models [3-5] to reduce these biases. However, previous studies were based on either city- or county-level data [4, 5] or simulated background data [3]. Some studies [2, 5] examined detection performance using single-day signals injection that may not represent real epidemic situations.

In this study, we used authentic daily syndrome counts of gastrointestinal (GI) time series at facility level as baseline data and added realistic data effects as target signals of disease outbreaks. We compared the detection performance of alerting algorithms on the basis of regression models and adaptive control-chart methods according to the use of total visits and management of day-of-week effects. We compared the sensitivity and timeliness of alerting to recommend methods for monitoring these data types.

2. Methods

2.1 Baseline data

The study data were from the CDC BioSense syndromic surveillance system. For background data, we used time series of daily syndrome counts at the facility level from the Department of Veterans Affairs (VA). Records were classified on the basis of ICD-9 codes from the Electronic International Classification of Diseases, 9th Revision codes using established Biosense syndrome groupings [1]. Study data were restricted to facilities that reported every day from Jan. 2010 through May 2011 and with median daily gastrointestinal (GI)[1] syndrome visit counts ≥3 over the entire study period. Many VA facilities had low outpatient visits on holidays. Therefore, the 14 federal holidays in the report period were recoded as weekend days.

We used a “sliding” baseline of 56 days to reflect the recent data behavior, with a 2-day buffer so that each test date was two days after the end of the baseline. The purpose of the 2-day buffer was to avoid contamination of the baseline data with a potential early phase of an outbreak [5]. The first baseline period began on 1/1/2010, and successive test dates went from 2/28/2010 through 5/31/2011.

With these restrictions of consistent and non-sparse reporting, the time series used for this study were from 59 facilities in 37 states. There were large variations in median count of outpatient visits among different facilities and of syndromes. We compared the sensitivity and timeliness of different surveillance detection methods applied to two scale categories of daily GI visit count time series, one category with median count levels below 10 and the second category with median levels ≥ 10.

2.2 Control-chart-based algorithms

Method 1: The count-based C2 was one of three control-chart-based algorithms tested. This method was equivalent to the C2 method developed for the Early Aberration Reporting System (EARS) [6] with a 56-day baseline. For each facility, the general form of the C2 test statistic is

$$C2 = \frac{(n_t - \mu)}{SD}$$

where n_t is the count for a syndrome on a test day (t), μ and SD are the arithmetic mean and the standard deviation of values in the sliding baseline.

Method 2: Count-based CuSUM has the same calculations of μ and SD as the Count-based C2 algorithm. While the C2 Count's decision rule relies on using one observation at a time, the CuSUM statistic incorporates information using past observations. The CuSUM statistics CS_t is calculated recursively as

$$CS_t = max\{0, CS_{t-1} + [(n_t - (\mu - k*SD))SD]\ \}$$

where k =0.5 is the magnitude of the shift to be detected in SD units above the mean, CS_t is the current CuSUM calculation, and CS_{t-1} is the previous CuSUM calculation [3, 6].

Method 3: Proportion-based C2 is similar to Count-based C2 except that it uses proportion (P_i) to calculate the arithmetic mean (μ) and standard deviation (SD) of index day where P_i=100*n_i/N_i, and n_i is the GI syndrome count, N_i is the total number of visits on baseline day i.

Method 4: The adjusted-baseline C2 is a rate-based version of the C2 algorithm [2]. In this approach, E_t is the expected number of GI syndromic visits for test day (t) and is calculated as

$$Et = Nt \times \frac{\sum_{i=1}^{56} n_i}{\sum_{i=1}^{56} N_i}$$

where Nt is the number of total visits for the test day. Note that E_t will differ from μ in C2 Proportion because it is adjusted by the total visits over the entire baseline period. This expected value is then used in the `adjusted' C2 test statistic

$$C2' = \frac{(n_t - Et)}{SD'}$$

where SD' is the adjusted standard deviation over baseline days estimated as

$$SD' = \frac{\sum_{i=1}^{56} \left| n_i - N_i \times \frac{\sum_{j=1}^{56} n_j}{\sum_{j=1}^{56} N_j} \right|}{56}$$

where n_j is the GI syndrome count and N_j is the total number of visits on baseline day *j*.

Method 5: The Adjusted-baseline CuSUM uses the same adjustment of total visits in calculating expected value (E_t) and standard deviation (*SD'*) as Method 4. However, it uses CuSUM statistic to incorporate information from past observations. Hence the Adjusted CuSUM statistic is

$$CS'_t = max\ \{0,\ CS'_{t-1} + [(n_t - (E_t - k*SD'))/SD']\ \}$$

For GI-related visit counts, the VA outpatient time series showed distinctly lower visit counts on weekends and holidays than on weekdays. We stratified the baseline days used into weekdays and weekend/holidays for the baseline mean and standard deviation calculations in the above methods. The 56-weekdays/weekends-stratified baseline days contained ~40 weekdays and ~16 weekend days. Each method was tested with and without this stratification.

2.3 Regression models

Two regression models were adapted from the tested models in a previous study [5]. The models included indicator variables to capture day-of-week effect and indicator variables for each 14-day interval (*Bi-week*) to adjust for seasonality. Both of the models also controlled for total daily visits.

<u>Method 6</u>: Linear regression model (Linear Reg) is the predicted value adjusted by using total visits as a linear covariate

$E_t = \beta_0 + \beta_1 * Monday + \ldots + \beta_6 * Saturday + \beta_7 * BiWeek_2 + \ldots + \beta_9 * BiWeek_4 + \beta_{10} * N \mid$ *distribution=normal, link=identity*

where Sunday is the reference day, BiWeek refers to a 2-week interval, and the reference 2-week interval is the first one, BiWeek_1.

<u>Method 7</u>: Poisson regression model (Poisson Reg) is the predicted value adjusted by using log form of total visits as offset.

$E_t = \beta_0 + \beta_1 * Monday + \ldots + \beta_6 * Saturday + \beta_7 * BiWeek_2 + \ldots + \beta_9 * BiWeek_4 \mid$ *distribution=Poisson, link=log, offset=log (N)*

The regression models were run separately for each facility, with the expected value for each test day predicted from the regression model from the previous 56 days of data, with a 2-day buffer. The *SD* of the expected value was calculated using the equation

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

where the E_i is the model expected number of GI syndrome count for a given day (i) in the 56-day baseline period. The detection statistic was computed by dividing the forecast residuals by this SD.

As for the chart-based methods in Section 2.2, both regression models were tested with and without weekend/weekday stratification. To avoid division by small numbers and thus unreasonably high statistic values, we used a minimum standard deviation of 0.2 for proportion methods and 1.0 for all other methods.

2.4 Outbreak signal simulation

In order to evaluate the performance of different aberration detection methods on real background data, we simulated outbreak signals. We generated the set of incubation periods with a set of random lognormal draws and rounded each to the nearest day. The number of cases to add for each day was then the number of draws rounded to that day [11].

2.5 Method evaluation

We created evaluation datasets by injecting the simulated outbreak signals successively into the GI syndrome time series for each facility, with the outbreak signals starting on March 1, 2010. Because the durations of signals ranged from 3 to 10 days, injection could begin any day of the week. We applied the test methods to the evaluation datasets for each facility and calculated the sensitivity and timeliness for detecting injected signals using four steps: (1) running each method to estimate expected value, SD, and test statistics for each facility day; (2) calculating an alerting threshold as the 99th percentile value for the test statistic; (3) calculating the test statistics for the test day using the time series with injects and recording as detections of the inject dates when the algorithm value exceeded the threshold; and (4) computing sensitivity and alerting timeliness on the basis of the recorded detections dates for each method.

Sensitivity was defined as the ratio of number of signals detected before the event peak to the total number of injected signals. Alerting timeliness was calculated as the delay until detection, which is the number of days from the start of injection to the first algorithm alert not later than the peak day of the injected signal. If there was no alert or an alert occurred after the peak day, the timeliness was set as 1 + the peak day on the basis of the lognormal distribution for the facility. For example, if the peak inject for a facility was on the third signal day, then a method alerting on the second day would be assigned a delay of 2, while a method alerting on the peak day (or not at all) was assigned a delay of 4.

3. Results

3.1 Descriptive data

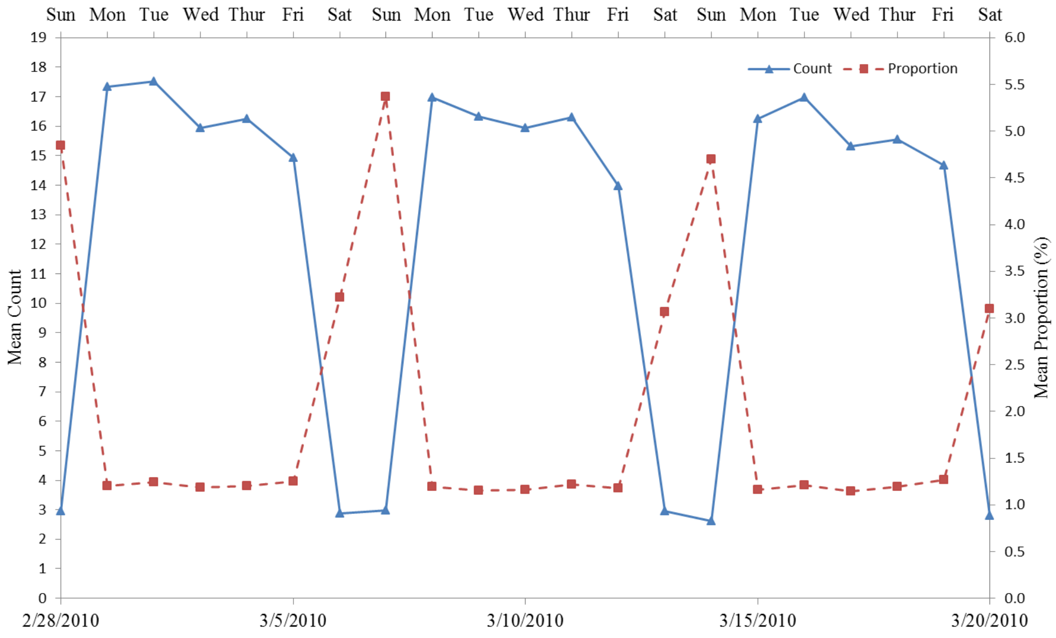
In the 59 study facilities, the mean of median daily total visits was 1,064 and mean of daily median GI count was 10.8 (Table 1). There were 18 and 41 facilities in median level of 3-9 and ≥10 category, respectively.

Table1. Summary of facility-level data by median of GI daily count level

Median of GI Count Level	Number of Facilities	Mean of Daily GI Count Medians	Mean of Daily Total Visit Medians
3-9	18	6.8	790
≥10	41	15.6	1531
Total	59	10.8	1064

The VA time series for daily GI syndrome counts showed a strong day-of-week effect. Figure 1 shows distribution of 59-facility's mean daily GI syndrome visit from Feb. 28 through March 20, 2010. The mean daily GI syndrome visit counts was higher (>13) on weekdays and lower (<4) on weekends/holidays. On the other hand, the mean daily proportion (percentage of a syndrome counts over the total visits) was lower (<1.5) on weekdays and higher (>3.0) on weekends/holidays (Figure 1). During week days, the mean daily GI syndrome counts were higher on Monday and Tuesday than the other days. On weekends, the mean of percentage of syndrome counts over the total visits was higher on Sunday than Saturday (Figure 1).

Figure 1. Distribution of mean GI syndrome counts and proportions (100*count/total visits) by day of week



3.2 Signal injection and detection

We tested 3,714 injected signals of GI syndrome. The mean duration of outbreak signal was 6.7 days (range 3 to 10) with the mean peak day of 2.4 days (range 1 to 5). The mean peak day value of injected signals ranged from 3 to 23 cases.

Figure 2 illustrates how a multiday signal was injected onto GI syndrome data from April 3 to 30, 2010, in one facility. Duration of the injected signal is 6 days. It peaked on the third day and the peak GI count was 21 (Figure 2, upper right). The signal was added onto the background data (solid curve) beginning on April 25 and ending on April 30. With 48 background cases, the total cases (x symbols) on April 27 (peak day) became 69 (=21+48) (Figure 2, lower). The figure shows the Poisson Reg model predicted value (triangle symbols) based on the 56-day sliding baseline (with a 2-day buffer) with weekday/weekend-holiday stratification.

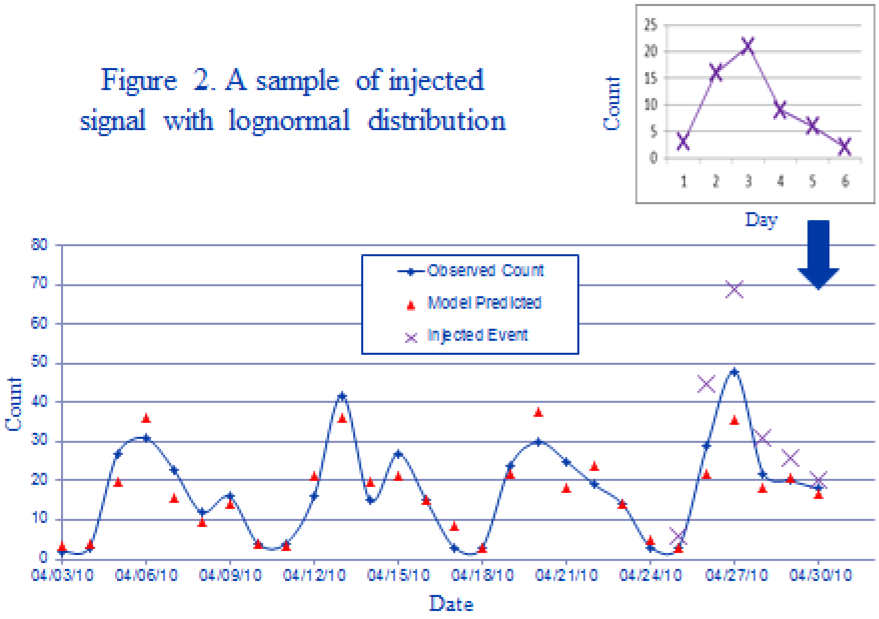


Figure 2. A sample of injected signal with lognormal distribution

3.3 Signal sensitivity

Sensitivity was compared among the different alerting methods using a detection threshold derived for a background alert rate of 1 threshold crossing per 100 days for the two time series scale categories, median <10 and median ≥10 (Table 2).

Without weekday/weekend stratification, at the 3-9 median level, the Adjusted-baseline C2 and Adjusted-baseline CuSUM methods outperform all other methods. At mean level ≥10, the two regression models have higher sensitivities than all control-chart-based methods. Weekday/weekend stratification increases sensitivities more than 10% for all methods. Proportion-based C2 has the lowest sensitivities at both mean levels. At the mean level 3-9, the Adjusted-baseline C2, Adjusted-baseline CuSUM, and two regression models perform better than three other control-chart based methods. At mean level ≥10, the two regression models have higher sensitivities than all control-chart-based methods. Poisson Reg consistently performs better than Linear Reg.

Table 2. Sensitivity to detect multi-day signal at 1% background alert rate

Wkday/ Wkend Strat.	Mean Level	Count-based C2	Count-based CuSUM	Proportion-based C2	Adjusted-baseline C2	Adjusted-baseline CuSUM	Linear Reg	Poisson Reg
No	3-9	26.1	27.7	34.3	43.4	43	36.9	37.1
	≥10	27.8	27.9	40.4	47.2	44.7	48.4	48.5
Yes	3-9	56.6	54.7	44.7	60.3	59.4	59.4	61.0
	≥10	57.0	55.1	48.3	58.7	59.2	64.4	65.7

3.4 Alerting timeliness

Timeliness of alerting a signal was compared among the methods using a threshold derived for a background alert rate of 1 threshold crossing per 100 days. We summarized the mean number of days until alerting a signal stratifying by the median daily syndrome count level (i.e. 3-9 and ≥10) (Table 3). A smaller alerting delay indicates better timeliness. Without weekday/weekend stratification, at the mean level of 3 to 9, Adjusted-baseline CuSUM method appears to have the best timeliness, though mean differences in Table 3 vary by less than a day. At the mean level ≥10, Adjusted-baseline CuSUM, Linear Reg, and Poisson Reg perform better than other methods. Weekday/weekend stratification shortened detection delays for all methods. Proportion-based C2 has the longest delay at both median levels. At the median level of 3 to 9, it appears that Adjusted-baseline CuSUM has the best timeliness. At median level ≥10, Poisson Reg performs better than all other methods.

Table 3. Timeliness (Mean delay days) to detect multi-day signal at 1% background alert rate

Wkday/ Wkend Strat.	Mean Level	Count-based C2	Count-based CuSUM	Proportion-based C2	Adjusted-baseline C2	Adjusted-baseline CuSUM	Linear Reg	Poisson Reg
No	3-9	3.15	3.00	2.94	3.09	2.88	3.02	3.02
	≥10	3.29	3.03	3.04	3.24	2.99	3.00	3.00
Yes	3-9	2.68	2.88	2.66	2.68	2.62	2.66	2.65
	≥10	2.78	2.95	2.79	2.78	2.72	2.69	2.66

4. Discussion

The purpose of this study was to identify the most robust and widely applicable aberration detection methods for an automated national biosurveillance system. By our study using daily facility-level VA visit counts of GI-related syndrome visits, we found that total-visit baseline adjustments can avoid alerting bias by removing data effects unrelated to outbreaks. Monitoring counts with adjusted baseline values gives superior detection performance to monitoring proportions. However, total-visit baseline adjustments do not remove all weekend effects. Weekday/weekend stratification improves performance for all methods. Poisson regression modeling yielded top performance for time series with daily median count ≥10. For sparser time series linear regression (e.g., the median daily count of 3 to 9), adjusted baseline C2 and adjusted baseline CuSUM are both effective.

We used real data from the BioSense national biosurveillance system as background data. Other studies [3, 6] based on simulated background data are typically too “clean” to represent authentic nonstationary data. The VA facilities whose data were included in our study were located in 37 U.S. states. Our analyses on facility-level data is of interest to many users, while others [4, 5] have used county- or city-level data and may find better detection performance using other methods at those levels. Several studies used longer sliding baselines (e.g., 56 days or 2 years) for regression models [3-5] than for control-chart-based methods; one might be tempted to attribute the regression-based methods’ superior performance to the additional baseline information. In our study we used the same sliding baseline length for all methods.

The use of multi-day signals allowed us to compare the aberration detection methods for timeliness in addition to sensitivity. We randomly simulated signals for each facility on the basis of random draws from a lognormal distribution. This process allowed us to test the algorithms with signals of varying duration and amplitude. Using the chosen lognormal parameters, we set the total number of injected cases for each signal to obtain a theoretical peak-day injected count of twice the standard deviation of the background data to compare the method performance on target signals of modest size [4].

Like other studies [2, 4, 5], we set the alert rates for each algorithm empirically, by applying the algorithms to each syndrome for each facility and found the threshold that would yield approximately 1 alert every 100 days. Some prior studies have tended to use rates between 3 to 5 alerts every 100 days, which may be too high for routine surveillance and could desensitize users, leading to reduced likelihood of following up on alerts [4]. Our results showed that for the same background alert rate, thresholds differ based on data scale, data pattern, and methods applied. These findings suggest that thresholds should be adapted in practice for data-stream types and alerting methods.

There were several limitations of this study. First, we included facilities with a median baseline daily syndrome count ≥ 3. Therefore, our results might not apply to data from facilities with sparse or occasional reporting. Second, all input time series were composed of daily VA hospital outpatient visit counts, so the time series reflect the coding and classification conventions at those facilities. Series characteristics that may affect algorithm performance may differ in data from other medical systems [4]. Third, the three commonly monitored syndrome groups used for the study data do not represent all practical classifications. Data derived for other syndromes may differ in seasonality and other systematic behaviors from the study series. Our results should not be considered representative of all monitored time series and practical signal types.

5. References

1. Tokars, J.I., et al., *Summary of data reported to CDC's national automated biosurveillance system, 2008.* BMC Med Inform Decis Mak, 2010. 10: p. 30.
2. Tokars, J.I., et al., *Enhancing time-series detection algorithms for automated biosurveillance.* Emerg Infect Dis, 2009. 15(4): p. 533-9.
3. Fricker, R.D., Jr., B.L. Hegler, and D.A. Dunfee, *Comparing syndromic surveillance detection methods: EARS' versus a CUSUM-based methodology.* Stat Med, 2008. 27(17): p. 3407-29.

4. Jackson, M.L., et al., *A simulation study comparing aberration detection algorithms for syndromic surveillance.* BMC Med Inform Decis Mak, 2007. 7: p. 6.

5. Xing, J., H. Burkom, and J. Tokars, *Method selection and adaptation for distributed monitoring of infectious diseases for syndromic surveillance.* J Biomed Inform, 2011. 44(6): p. 1093-101.

6. Hutwagner, L., et al., *Comparing aberration detection methods with simulated data.* Emerg Infect Dis, 2005. 11(2): p. 314-6.

7. Armenian, H.K. and A.M. Lilienfeld, *Incubation period of disease.* Epidemiol Rev, 1983. 5: p. 1-15.

8. Philippe, P., *Sartwell's incubation period model revisited in the light of dynamic modeling.* J Clin Epidemiol, 1994. 47(4): p. 419-33.

9. Sartwell, P.E., *The distribution of incubation periods of infectious disease. 1949.* Am J Epidemiol, 1995. 141(5): p. 386-94; discussion 385.

10. Detrick, F., *Medical Management of Biological Casualties.* U.S. Army Medical Research Institute of Infectious Diseases, Sept. 2000.

11. Johnson, N.L.K., Samuel; Balakrishnan, N. , *Continuous univariate distributions.* . Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, 1994. 1: p. "14: Lognormal Distributions",.

12. Pavlin, J.A., et al., *Combining surveillance systems: effective merging of U.S. Veteran and military health data.* PLoS One, 2013. 8(12): p. e84077.