SAM SCHAEFER

Department of Mathematical Sciences
Montana State University

May 3, 2013

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

SAM SCHAEFER

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

5/1/2013
Date

Steve
YOUR ADVISOR'S NAME HERE
Writing Project Advisor

5/1/2013
Date

Megan D. Higgs
Writing Projects Coordinator

Biosurveillance: Case-Detection Algorithms and Forecasting Methods

Sam Schaefer

May 3, 2013

Abstract

Biosurveillance is the science of disease outbreak detection. Early detection provides health officials additional time to fight the spread of disease. Throughout history, humans have continually developed new methods to fight the spread of disease, beginning with disease diagnoses. Once cases are diagnosed, they can be followed prospectively and compared to an epidemic threshold, a mark indicating an alarm should be raised signaling the start of a disease outbreak. This paper examines a common structure used in case-detection diagnostic systems as well as classical time series forecasting methods used to establish the alarm threshold level. In addition, forecasting methods adopted from statistical quality control and their application to surveillance forecasting are also explored.

Introduction

Biosurveillance is the science of real-time disease outbreak detection in people, plants, and animals. This science applies to both natural and man-made epidemics. Man-made epidemics are classified as acts of bio-terrorism. Through the use of efficient biosurveillance systems, health specialists hope to be able to detect a disease outbreak as soon as possible. This would allow ample time to provide necessary medications and resources to slow the spread of the disease. While traditional methods of biosurveillance were concerned with studying outbreaks retrospectively, current systems attempt to detect outbreaks as early as possible through the use of syndromic data, which are early presentations of an illness. This paper explores the history of disease outbreaks and biosurveillance as well as some of the case-detection and forecasting methods employed to detect and slow the onslaught of these outbreaks.

Disease Outbreak History

Interactions between humans and microbial organisms have been present throughout history and many have been beneficial. For instance, it is believed that mitochondria, the source of energy that fuels our cell processes, evolved from bacteria in early stages of evolution. These interactions are not always beneficial however. "Human populations have been battling these unseen living organisms throughout the course of history and, in many instances, losing" (Wagner, Moore, Aryel, 2006, p. 13). Some instances of these outbreaks are, but are not limited to, the Black Plague, the 1918 influenza pandemic, the Lyme Disease outbreak of 1975 in Lyme, Connecticut, the Soviet Union anthrax outbreak of 1979, the United States AIDS epidemic of 1981, and the SARS epidemic of 2003. These outbreaks are not limited to natural causes however. The anthrax attacks of 2001, which occurred only one week after the September 11 attacks, made evident the possibility of a large scale bio-terrorism act. This led to increased efforts in biosurveillance in the past decade.

Biosurveillance History to Present

Disease reporting is thought to have evolved into its current form from 1949 to 1970. Much of the credit is given to Alexander Langmuir, the creator of the Epidemic Intelligence Service, a program within the American Centers for Disease Control and Prevention (CDC). Langmuir proved to be influential in the drive for disease reporting. Currently, the lead agency related to biosurveillance at the federal level is the CDC. In addition to operating at the federal level, the CDC is responsible for the collection, analysis, and dissemination of disease occurrence and mortality data to both state and local health departments. The availability of these data gave agencies the ability to study and diagnose past occurrences of epidemics. Until the early 1970's, the study of disease occurrences was reserved for physicians. This specific discipline soon shifted to epidemiologists, individuals with extensive skills in questionnaire designs and epidemiological methods. Furthermore, they typically had extensive training in statistical analysis. Unlike physicians, this allowed epidemiologists to examine and interpret large quantities of case data, as opposed to studying individual patients.

The acquisition of data needed to characterize outbreaks begins at the case detection stage. This stage is concerned with noticing the existence of a disease within a single individual. The entities most often associated with case detection are physicians, veterinarians, nurse practitioners, pathologists, laboratories, and surveillance systems. Case detection by medical specialists and laboratories is simply a product of routine operation. In an attempt to thwart attempted bioterrorism attacks in addition to natural outbreaks, the case detection method of drop-in surveillance has been employed (Mandi & Paola). This method examines physicians' reports on numerous patients in the weeks leading up to large events, such as the Olympics, and records whether the patients meet various syndromes of interest. The hope is that in the weeks leading up to a large event a baseline case rate can be established. Once an event is underway, disease cases can be compared to this baseline rate on a daily basis to examine possible indications of an outbreak.

Finally, advances in computing power have allowed case detection through computers. Case detection through computers is currently employed in an attempt to detect syndromes. The majority of infectious diseases of interest typically present an initial set of syndromes prior to their manifestation. These computer based systems monitor for syndromes associated with diarrhea, respiratory illnesses, hemorrhagic, and influenza-like conditions. The techniques used to detect and classify these syndromes will be explored later.

Once systems were established that could provide case data over long periods of time from one location to another, epidemiologists were able to study these data to determine the existence of an outbreak. The classification of outbreaks was performed retrospectively, typically weeks or months after a given outbreak had occurred. While these early analyses provided insight into possible causes of outbreaks, they were useless for early outbreak detection. However, as surveillance databases provided up-to-date case data, outbreak detection methods were developed to predict outbreaks before large-scale infestations were reached. Early detection would allow ample time for the necessary resources to be dispersed to the individuals most susceptible to a specific disease. These methods are currently employed and are based on previous days' case counts. Particular methods related to these practices will be discussed later.

In recent years surveillance systems have shifted their focus from not only analyzing case counts over time, but by also exploring syndromic data. Syndromic data are data associated with the earliest indicators of disease presence. Examples of such data are over-the-counter (OTC) drug sales, absenteeism rates, online health information rates, phone calls to nurse hotlines, and many more. To use these data, there is an underlying assumption that individuals will attempt to self-treat before visiting a medical professional (Burkom & Shmueli, 2010). The hope is that a disease outbreak will manifest itself through an anomaly in syndromic data, compared to background behavior of these syndromic data. One of the large and ongoing problems associated with using syndromic data is defining the background behavior. Once this background behavior is defined however, a threshold level can be created that dictates whether an outbreak alarm is signaled. This threshold is typically between 2 and 3.5 standard deviations of a model's prediction, which is developed from background behavior data.

Case-Detection Algorithms

Prior to predicting or detecting an outbreak, individual disease cases must be identified. Thus, case detection is an essential part of biosurveillance. Following the anthrax attacks of 2001, efforts increased to create a large-scale disease detection system. In a perfect world, each day every individual could enter their current health information into a system, and the computer could predict, with associated probabilities the likelihood that the individual had a specific disease. This would allow up-to-date data on all individuals, making it easier to detect the presence of an outbreak. While this is obviously not feasible, the implementation of these systems at a smaller scale and their associated automated methods can prove to be useful.

The programs used to calculate various case-specific probabilities are known as diagnostic expert systems. These systems, like physicians, use patients' symptoms to provide a specific diagnosis.

Typically, patients' symptons are entered into a system by a physician or an assistant. Stored within these systems are data on disease prevalences as well as sensitivity and specificity estimates for disease tests. Disease prevalence is defined as the proportion of a population found to have a certain condition. The sensitivity of a test is the probability that the test will detect the disease when the disease is present. In addition, the specificity of a test is the probability that a test will detect that a disease is absent in situations when an individual lacks the disease of interest. Thus the corresponding equations for sensitivity and specificity are as follows.

$$\text{sensitivity} = P(T^+|D^+) \tag{1}$$

$$\text{specificity} = P(T^-|D^-) \tag{2}$$

In Equations 1 and 2, $T+$ and $T-$ represent the result of the test while D^+ and D^- represent the presence and absence of a disease, respectively. Given disease prevalence and associated sensitivity and specificity of tests for a disease, these programs proceed to generate a differential diagnosis for a sick patient. This differential diagnosis is simply a list of diseases most likely to be responsible for a specific patient's symptoms. The differential diagnosis uses the corresponding symptoms to calculate a posterior probability for every disease that is likely responsible for the symptoms. These diagnoses are often computed using Bayes' rule. To begin, the prior probability for a disease is defined as the disease prevalence. These will be represented as P(Disease)=P(D+). Given specific symptoms, also called findings, a posterior probability can be computed for a specific disease. These case-detection systems use Bayes' rule to compute the posterior probability of a specific disease. The example below demonstrates the basic principles of many case-detection systems. Specifically, the Bovine-Syndromic Surveillance System (BOSSS), which is a web-based disease-reporting surveillance system for cattle, uses an odds-likelihood form of Bayes' rule. This odds-likelihood form of Bayes' rule uses the sensitivity and specificity of given tests to compute the posterior probability of a given disease. Before exploring the odds-likelihood form of Bayes' rule used by BOSSS, Bayes' rule and its application to case-detections will be explored.

To begin, suppose there are several diseases of interest thought to be the cause of a particular set of findings (symptoms). These diseases can be represented as $D_1, D_2, ..., D_n$ where D_1 represents disease 1, D_2 represents disease 2, and D_n represents a disease free state. To calculate the posterior probability of disease 1 given a set of findings F, Bayes' rule can be used as shown in Equation 3.

$$P(D_1^+|F) = \frac{P(F|D_1^+) * P(D_1^+)}{P(F|D_1^+) * P(D_1^+) + P(F|D_1^-) * P(D_1^-)} \tag{3}$$

The data used for the example below are in Tables 1 and 2. These data and application to these methods are courtesy of the text "Handbook of Biosurveillance" (Wagner, Moore, Aryel, 2006).

Prior Probabilities and Odds of FMD and MCD

Disease	P(Disease)	Odds(Disease)
Foot and Mouth Disease (FMD)	0.001	0.001001
Mad Cow Disease (MCD)	0.001	0.001001

Table 1

Conditional Probabilities for FMD and MCD

Finding	Disease	p(Finding|Disease)
Drooling of saliva present	FMD present	0.95
Drooling of saliva present	FMD absent	0.05
Drooling of saliva present	MCD present	0.001
Drooling of saliva present	MCD absent	0.05
More than one animal affected	FMD present	0.95
More than one animal affected	FMD absent	0.2
More than one animal affected	MCD present	0.001
More than one animal affected	MCD absent	0.2

Table 2

Suppose that a single cow is observed drooling saliva and it is also observed that a nearby cow is also sick. Given these findings, the posterior probability of the cow having foot and mouth disease can be computed. The steps used to calculate this posterior probability are shown below. First, the quantity of interest is:

$$\text{P(FMD}^+\text{|drooling of saliva present, other sick cow detected)}$$

$$P(FMD^+|f_1, f_2, ..., f_m) = \frac{\prod_{k=1}^{m} P(f_k|FMD^+) * P(FMD^+)}{\prod_{k=1}^{m} P(f_k|FMD^+) * P(FMD^+) + \prod_{k=1}^{m} P(f_k|FMD^-) * P(FMD^-)} \quad (5)$$

Using Equation 5 above, as well as the information provided in Tables 1 and 2, the posterior probability of foot and mouth disease can be computed given the findings of drooling and another sick cow present.

$$P(FMD^+|d, sc) =$$

$$= \frac{P(d|FMD^+) * P(sc|FMD^+) * P(FMD^+)}{P(d|FMD^+) * P(sc|FMD^+) * P(FMD^+) + P(d|FMD^-) * P(sc|FMD^-) * P(FMD^-)}$$

$$= \frac{0.95 * 0.95 * 0.001}{0.95 * 0.95 * 0.001 + 0.05 * 0.2 * 0.999} = 0.083$$

Similarily, given these two findings, the posterior probability for mad cow disease can also be computed. Once again, the data from Tables 1 and 2 will be used for this calculation.

$$P(MCD^+|d, sc) =$$

$$= \frac{P(d|MCD^+) * P(sc|MCD^+) * P(MCD^+)}{P(d|MCD^+) * P(sc|MCD^+) * P(MCD^+) + P(d|MCD^-) * P(sc|MCD^-) * P(MCD^-)}$$

$$= \frac{0.001 * 0.001 * 0.001}{0.001 * 0.001 * 0.001 + 0.05 * 0.2 * 0.999} = 1 * 10^{-7}$$

Therefore, the posterior probabilities of foot and mouth disease and mad cow disease, given the findings, are 0.083 and $1 * 10^{-7}$ respectively. While this method works well, using the odds-likelihood method is much less cumbersome.

To describe the odds-likelihood form of Bayes' rule adapted by BOSSS, the definition of odds must first be defined. The odds of an event are simply defined as:

$$odds = \frac{p}{1-p} \quad (6)$$

In Equation 6, $p = P(\text{Event occurs})$ and the quantity $1 - p = P(\text{Event does not occur})$. Given this

Using d to represent drooling and sc to represent another sick cow, Equation 3 can be written as:

$$P(FMD^{+}|d, sc) =$$

$$\frac{P(d, sc|FMD^{+}) * P(FMD^{+})}{P(d, sc|FMD^{+}) * P(FMD^{+}) + P(d, sc|FMD^{-}) * P(FMD^{-})}$$

To begin, the term in the numerator will be evaluated first. Note that for this computation as well as that used by the BOSSS odds-likelihood method, the assumption is made that all findings are independent of one another. In practice, the application of this assumption is known as a naive Bayes classifier. While this assumption seems overly simplistic, BOSSS's pilot studies showed that this system provided disease reports comparable to those created by veterinarians (Shepard, Toribio, Cameron, Thomson, Baldock. 2006). Therefore, for a set of m findings $f_1, f_2, ..., f_m$, the probability of observing them given FMD is as follows.

$$P(f_1, f_2, ..., f_m|FMD^{+}) = \prod_{k=1}^{m} P(f_k|FMD^{+}) \tag{4}$$

For this particular example, there are two findings, namely drooling and another sick cow. Therefore, in Equation 4 these findings represent f_1 and f_2. After substituting this quantity into the numerator, Equation 3 now becomes:

$$\mathrm{P}(FMD^{+}|d, sc) = \frac{P(d|FMD^{+}) * P(sc|FMD^{+}) * P(FMD^{+})}{P(d, sc|FMD^{+}) * P(FMD^{+}) + P(d, sc|FMD^{-}) * P(FMD^{-})}$$

Using the law of total probability along with application of Equation 4, the denominator equates to:

$$P(d, sc|FMD^{+}) * P(FMD^{+}) + P(d, sc|FMD^{-}) * P(FMD^{-}) =$$

$$P(d|FMD^{+}) * P(sc|FMD^{+}) * P(FMD^{+}) + P(d|FMD^{-}) * P(sc|FMD^{-}) * P(FMD^{-})$$

Therefore, for multiple findings, after combining the numerator and denominator terms from above, Equation 3 can be expressed as:

$$= \frac{P(\text{drooling saliva}|\text{FMD}^+)}{P(\text{drooling saliva}|\text{FMD}^-)} * \frac{P(\text{other sick cow}|\text{FMD}^+)}{P(\text{other sick cow}|\text{FMD}^-)} * Odds(FMD)$$

$$= \frac{0.95}{0.05} * \frac{0.95}{0.20} * 0.001001 = 0.09$$

Using the formula $probability = \frac{\text{odds}}{1+\text{odds}}$ the posterior odds of 0.09 equals a posterior probability of 0.083, the value obtained using the previous method.

In addition to the BOSSS diagnostic expert system, prominent systems for human diagnostics are the Iliad, DXplain, and Simulconsult systems. These systems perform on a nearly equivalent level to physicians, though are not used widely in today's medical practices as the data entering has proven to be extremely time-consuming. However, the utilization and efficiency of these systems for cattle continues to be explored.

Classical Time Series Forecasting Methods

Data on a multitude of disease counts are currently readily available for many health agencies. The counts are continually examined by surveillance systems in an attempt to detect an anomaly in the number of occurrences of a specific disease at any point in time. Through the use of these surveillance systems, health officials hope to detect a disease epidemic as soon as possible and provide the necessary resources to slow the spread of the disease at hand. The methods used will be explored using both real and simulated data.

Figure 1 below displays the Center for Disease Control and Prevention (CDC) weekly counts of influenza, from 2003 to the beginning of 2010 in the United States.

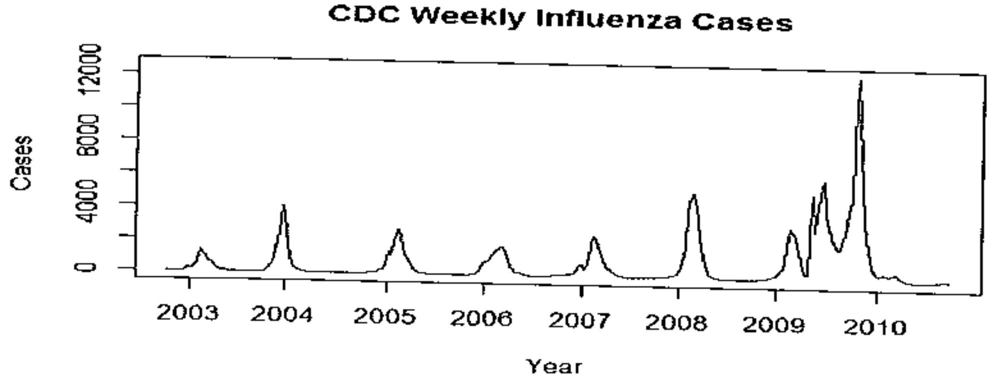

Figure 1: *Weekly influenza cases in the United States from 2003 to 2010.*

definition of odds, the odds-likelihood form of Bayes' rule is defined as:

$$Odds(D|f) = LR_{f|D} * Odds(D) \tag{7}$$

In Equation 7, $LR_{f|D}$ is the likelihood ratio, not to be confused with the likelihood-ratio test. The likelihood ratio defined above is expressed in terms of sensitivities and specificities. Two versions of this likelihood ratio exist, one for positive test results and one for negative test results. The likelihood ratio positive (LR^+) and likelihood ratio negative (LR^-) are defined as follows:

$$LR^+ = \frac{P(T^+|D+)}{P(T^+|D-)} = \frac{\text{sensitivity}}{1 - \text{specificity}} \tag{8}$$

$$LR^- = \frac{P(T^-|D+)}{P(T^-|D-)} = \frac{\text{1-sensitivity}}{\text{specificity}} \tag{9}$$

The derivation of Equation 7 is quite trivial. Starting with our definition of odds in Equation 6, $Odds(D|f)$ can be expressed as:

$$\begin{aligned} Odds(D|f) = \frac{P(D|f)}{P(D^C|f)} = \frac{\frac{P(D\text{and}F)}{P(F)}}{\frac{P(D^C\text{and}F)}{P(F)}} = \frac{P(D\text{and}F)}{P(D^C\text{and}F)} \\ = \frac{P(f|D) * P(D)}{P(f|D^C) * P(D^C)} = \frac{P(f|D)}{P(f|D^C)} * \frac{P(D)}{P(D^C)} \\ = LR_{f|D} * Odds(D) \end{aligned}$$

Using the odds-likelihood version of Bayes' rule will yield the same posterior probabilities as the previous method. Once again, the assumption is made that any findings are independent of one another. Similarly to the previous example, assume a cow is observed drooling saliva and there is another observed sick cow. Given these observations, the posterior odds for foot and mouth disease to calculate are Odds(FMD|drooling saliva, other sick cow). Since the cow was observed drooling saliva and another sick cow is present, the likelihood positive ratios are the appropriate ratios to use. This yields:

$$Odds(FMD|\text{drooling saliva, sick cow}) = LR^+_{\text{drooling saliva}|FMD} * LR^+_{\text{other sick cow}|FMD} * Odds(FMD)$$

The first forecasting method that will be discussed is the use of control charts to detect disease outbreaks. In biosurveillance, these control charts operate by creating an upper control limit (UCL). Using this method, outbreaks are defined when a disease count surpasses this UCL. To construct the UCL, the mean count and standard deviation from background time periods must be calculated. If $X_1, X_2, ...X_N$ represent counts from the background activity, then the background mean estimate $\hat{\mu}$ and standard deviation estimate $\hat{\sigma}$ can be calculated as shown below:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{10}$$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu})^2} \tag{11}$$

The underlying assumption used here is that the count data from the background activity follows a Normal distribution. The UCL is then defined as:

$$UCL = \hat{\mu} + k\hat{\sigma} \tag{12}$$

In Equation 12, k is typically equal to either 2 or 3. Given the assumption that the background activity follows a Normal distribution, setting k equal to 2 allows a 5% chance of mistakenly characterizing a count as an outbreak while k equal to 3 corresponds to a 1% chance. Using the data from Figure 1, $\hat{\mu}$ and $\hat{\sigma}$ were continually updated using the past N observations. Figure 2 displays the UCL plotted alongside the influenza data from Figure 1.

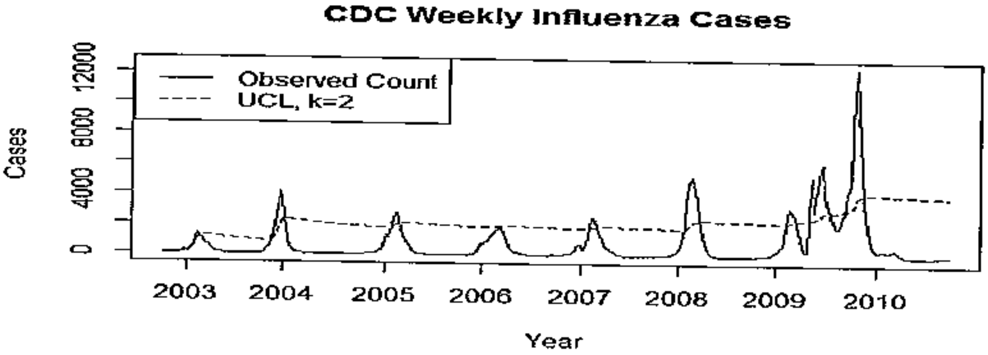

Figure 2: *Influenza case data plotted alongside control chart UCL using k=2.*

In Figure 2, the control chart method properly detects every outbreak. However, it is far too sensitive as it fails to account for seasonality that may be present in certain diseases, such as influenza. Furthermore, while it is common to observe many small false alarms, the differences between the observed counts and UCL during influenza season in Figure 2 are very large. One way to account for seasonality that may be present in a disease is to forecast counts using a select few of the past observations. This is accomplished by using an approach similar to the control chart method. First, a normal distribution is fit to the previous N observations. The estimated mean and standard deviation for this distribution are calculated using Equations 10 and 11. Thus, the predicted count value for any time t is defined as the average of the previous N observations, as shown below.

$$X_{t+1} = \frac{1}{N}(X_t + X_{t-1} + X_{t-2} + \ldots + X_{t-(N-1)})$$

Similarly to the control chart, a threshold value can be computed by adding k standard deviations to the forecasted count value using $\hat{\sigma}$ from the previous N observations . Using a 4-week moving average with k=2, Figure 3 displays the new threshold plotted alongside the count data from Figure 1.

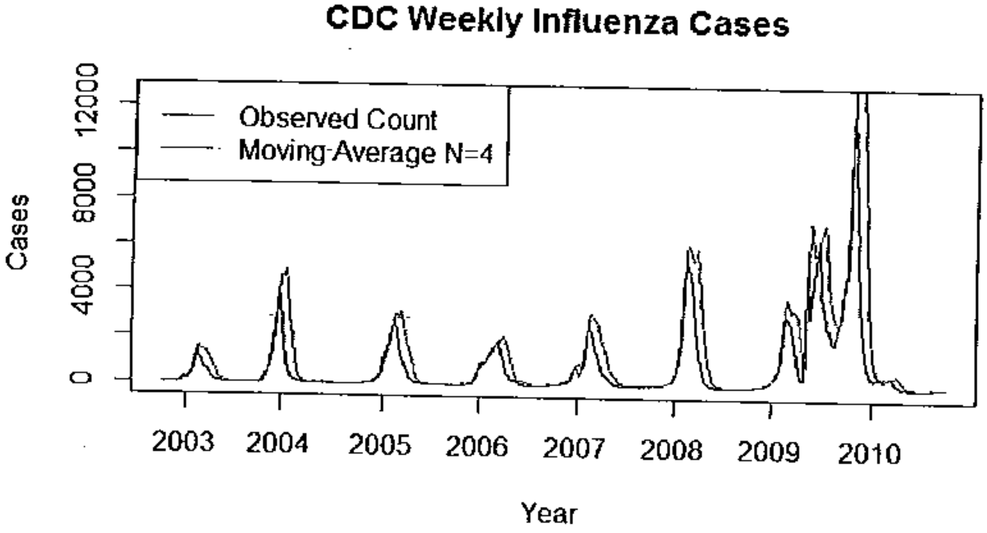

Figure 3: *4-week moving average threshold level.*

As Figure 3 above shows, the moving average threshold performs much better than the control chart threshold as it accounts for the seasonality nature of influenza. However, for these data there are still many instances in which the observed count is slightly above the threshold, which is not uncommon as noted earlier. Figure 4 below provides threshold levels for a two-week moving average, a six-week moving average, and an eight-week moving average in addition to the four-week moving average from Figure 3. All threshold values were created using an α=0.05 significance level.

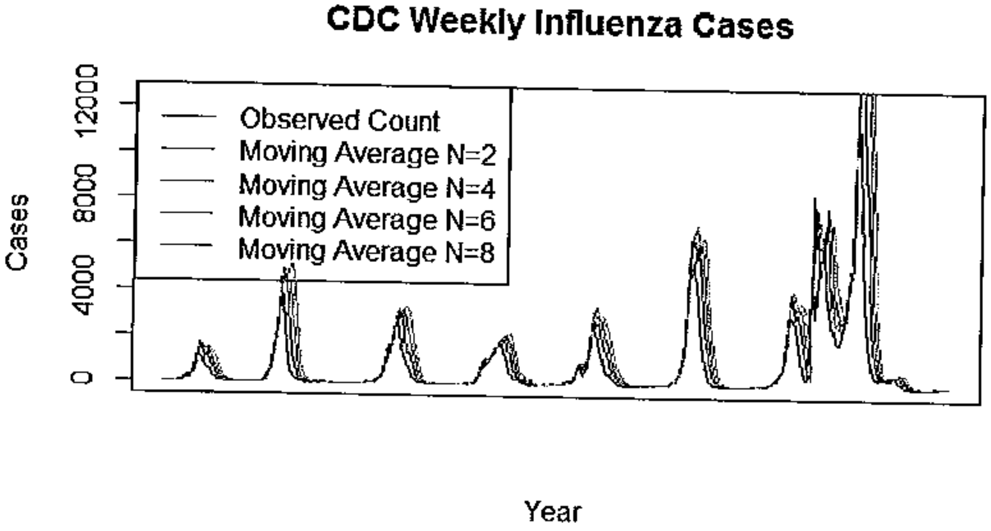

Figure 4: *2, 4, 6, and 8-week moving average threshold levels.*

As it is difficult to tell from the graph alone, the performances vary across the different values chosen for N. Table 3 below summarizes the results of the control chart and moving average forecasting methods.

Method	Weeks Above Threshold
UCL	61
Moving Average: N=2	108
Moving Average: N=4	98
Moving Average: N=6	107
Moving Average: N=8	107

Table 3

The table above suggests that the moving average forecasting methods would signal more alarms than the control chart method though the magnitude of the alarms raised for the control chart were much larger. As noted earlier, the distribution of the previous N-weeks' counts were assumed to come from a Normal distribution and parameter estimates were obtained using Equations 10 and 11. Instead of using Equations 10 and 11 to provide the estimates, it may make more sense to use maximum likelihood estimation to provide the estimates. Given today's widely available statistical software, this is quite simple to do. For each set of N observations preceding a given time, an estimate of μ and σ can be computed using MLE. Table 4 below summarizes the results for the control chart as well as the two estimation techniques for the moving average.

Method	Weeks Above Threshold
UCL	205
Moving Average: N=2	108
Moving Average: N=4	98
Moving Average: N=6	107
Moving Average: N=8	107
MLE Moving Average: N=4	98
MLE Moving Average: N=6	118
MLE Moving Average: N=8	115

Table 4

As Table 4 above shows, the MLE four-week moving average yielded the same amount of observed weeks above the threshold. These may not be the same weeks however. Furthermore, for the six and eight-week moving averages the MLE threshold was surpassed more than the original moving average method. To better handle the expected seasonality spikes, regression analysis with seasonal components is often used. To determine proper forecasting methods from one disease to another, simulation studies on numerous artificial disease-specific outbreaks are studied and compared. Methods with poor sensitivity are obviously unacceptable, as numerous outbreaks will not be detected. However, it is difficult to estimate the cost of numerous false alarms. Therefore, some appropriate middle ground must be chosen. These comparison techniques will not be discussed further in this paper.

To allow for easier comparison between these techniques and others, a simulated influenza time series is shown below with an induced outbreak beginning on the 59^{th} day. Also note that on day 31 there is a slight anomaly from the background behavior, though ideally an alarm should not be sounded here.

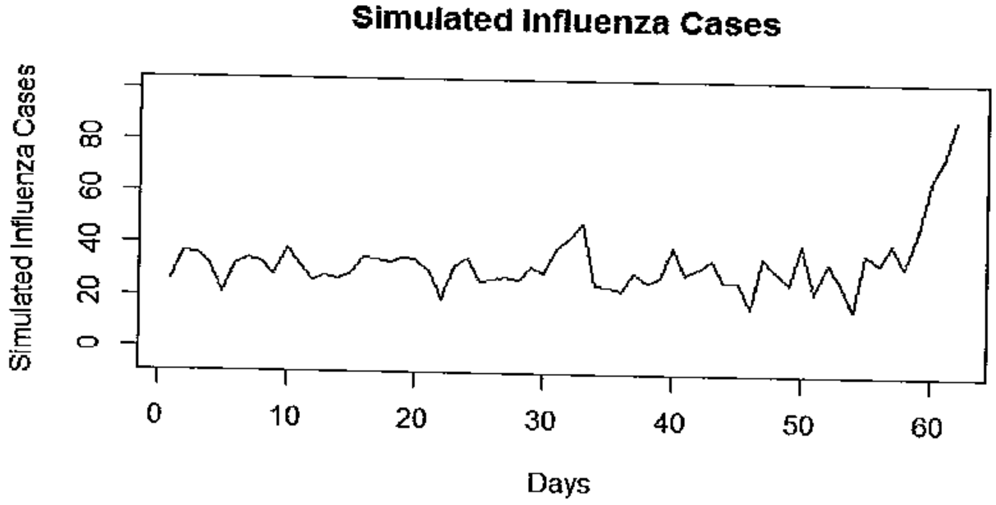

Figure 5: *Simulated influenza data with induced outbreak.*

Prior to examining new methods, the previous moving average techniques were employed here. Since these are daily data, four-day, one-week, and two-week moving averages were computed. These three moving averages' thresholds plotted alongside the data are shown below in Figure 6.

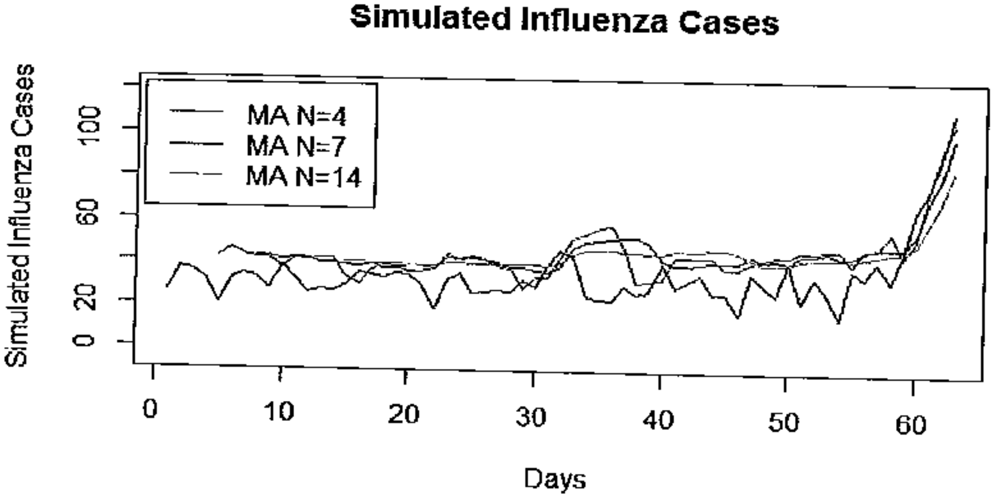

Figure 6: *4-day, 1-week, and 2-week moving average thresholds.*

Table 5 below provides a summary of results for the three moving average methods described above.

Method	Days Above Threshold	Days to Detect Outbreak
Moving Average: N=4	10	1
Moving Average: N=7	6	2
Moving Average: N=14	8	1

Table 5

From this one simulation the one-week moving average yields the least amount of false alarms. However, as a consequence it takes an additional day to detect the induced outbreak. These are the trade-offs that need to be considered when evaluating these methods. Two additional methods that are commonly used in disease forecasting are cumulative sum control charts (CUSUM) and exponentially weighted moving averages (EWMA).

The cumulative sum technique is a method used to detect small shifts in the mean of a process. The cumulative sum continually sums deviations from a desired reference value μ_0 until the deviations exceed some threshold level. As seen in the control charts and moving average examples, interest lies in observed counts exceeding an upper threshold level only. Here, the reference level μ_0 is defined as the in-control process mean. The definition of the CUSUM is described as follows.

$$C_i = X_i - \mu_0 + C_{i-1}$$

Above, X_i is the i^{th} observation and C_i is the i^{th} cumulative sum. Therefore, it then follows that the second cumulative sum up to the N^{th} cumulative sum can be expressed by:

$$C_2 = (X_2 - \mu_0) + (X_1 - \mu_0) = (X_2 - \mu_0) + C_1$$
$$\vdots$$
$$C_N = (X_N - \mu_0) + C_{N-1}$$

As previously noted, the only concern is if counts exceed an upper threshold. Therefore, the upper cumulative sum of interest is defined as:

$$C_i^+ = \max[0, X_i - (\mu_0 + k) + C_{i-1}^+] \tag{13}$$

In Equation 13, k is called the reference or allowable value and is defined as $k = \frac{|\mu_1 - \mu_0|}{2}$ where μ_1 is defined as the out-of-control process mean. The cumulative sum in Equation 13 is then compared to some decision interval H. Koshti (2011) references Montgomery (2001) that H should be equal to five multiplied by the in-control process standard deviation. Using this decision interval, alarms are signaled when C_i^+ reaches H. When this happens, the algorithm resets to zero and starts over. The plot below shows that the cumulative sum exceeds the decision threshold, H, on day 60.

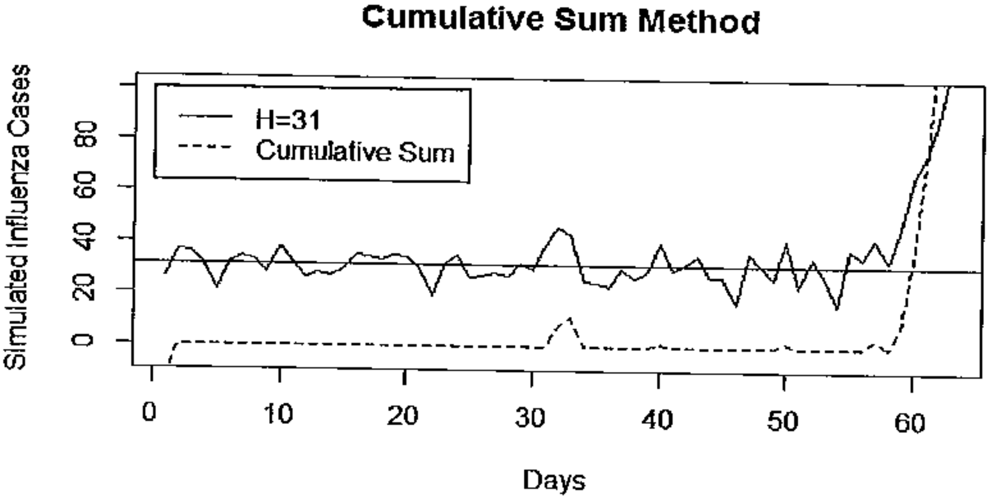

Figure 7: *Cumulative sum and decision threshold.*

For these simulated data, this method required an additional day to detect the induced outbreak. However, there was not a single false alarm leading up to the real outbreak. R Statistical Computing Software contains a surveillance package (Hohle, 2007) with a cumulative sum function *algo.cusum.* This function approximates the cumulative sum based on standardized observed counts. This function only accepts series at the weekly or yearly level. Therefore, the simulated time series used above was assumed to be weekly data when the *algo.cusum* function was employed. Figure 8 implies that for this one simulation, the *algo.cusum* function was slightly more sensitive than the algorithm used above in Figure 7. However, both CUSUM algorithms detected the induced spike at the end of the series.

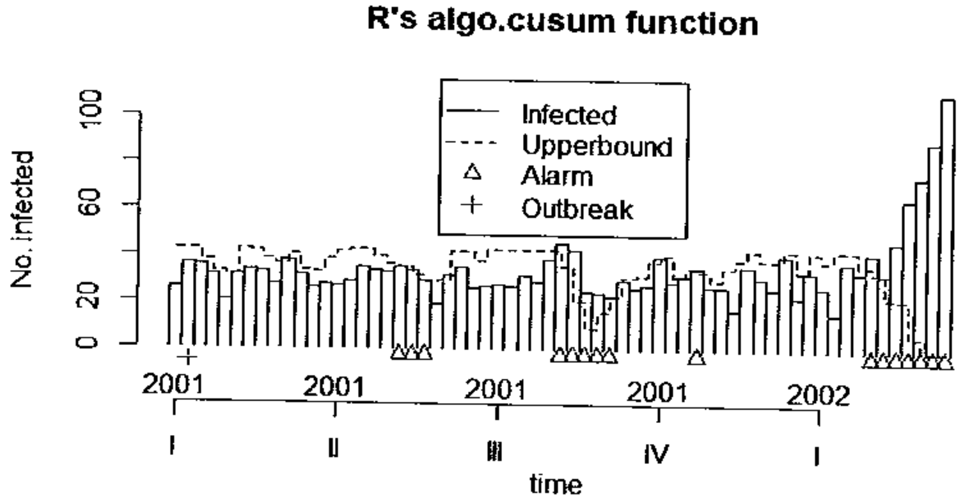

Figure 8: *R's algo.cusum results.*

The final forecast method to be explored has recently garnered attention in health surveillance, and was adopted from statistical quality control. Instead of giving equal weight to each observation, as seen in the moving averages, exponential weighting allows the weight given to a particular observation to decay as the observation becomes older. In general, exponential smoothing forecasts are simply weighted averages of past observations, where older observations are given less weight. Using this technique the EWMA statistic for some time t is defined by:

$$L_t = \alpha X_t + (1-\alpha)L_{t-1} \tag{14}$$

In Equation 14, α is a fixed smoothing coefficient between 0 and 1. Typically, α values between 0.1 and 0.3 are used (Burkom, Murphy, Shmueli, 2007). Expanded, Equation 14 is as follows:

$$L_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \ldots \tag{15}$$

Each individual L_t is then compared to some UCL. If L_t reaches the UCL an alarm is raised. Note that an initial value for L_t must be chosen. Often, an estimate of the in-control process mean or the first observed count is used (Does, Meulen, Vermat, 2008). Under the assumption that the $X_i's$

are independent random variables, the variance of the statistic, L_t, derived by Zhang (1998) is:

$$\sigma^2_{L_t} = \sigma^2(\frac{\alpha}{2-\alpha})[1-(1-\alpha)^{2t}] \tag{16}$$

In Equation 16, σ^2 is an estimate of the variance of the observations under outbreak-free conditions. Furthermore, depending on the size of α chosen, after the first few observations Equation 16 simplifies to:

$$\sigma^2_{L_t} = \sigma^2(\frac{\alpha}{2-\alpha}) \tag{17}$$

Using this variance, an upper control limit for the test statistic, L_t is computed using:

$$UCL = \mu_0 + k\sigma_{L_t}$$

Above, μ_0 represents the in-control process mean and a value of 3 for k is often suggested (Montgomery, 2001). Similarly to the cumulative sum approach, this method does not signal any false alarms and requires one day to detect the outbreak. This is shown below in Figure 9.

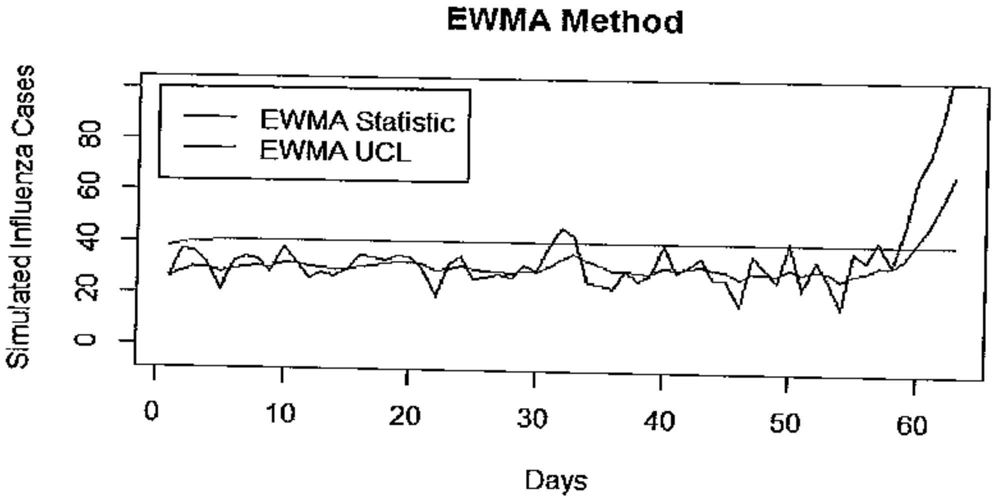

Figure 9: *EWMA statistic and EWMA UCL.*

The methods explored here are a small sample of techniques used to detect disease outbreaks. Furthermore, while these methods assumed a normal distribution to obtain alarm levels, it is not uncommon to create these levels using Poisson distributions. Table 6 summarizes the results of the forecast methods on this simulated outbreak.

Method	False Alarms	Days to Detect Outbreak
Moving Average: N=4	6	1
Moving Average: N=7	2	2
Moving Average: N=14	3	1
CUSUM	0	2
EWMA	0	2

Table 6

Both the CUSUM and EWMA methods required two days to detect the outbreak and signaled no false alarms. However, the one week moving average required only one day to detect the outbreak while signaling only two false alarms. It is precisely these trade-offs between false alarms and detection timeliness that need to be addressed when evaluating methods.

Conclusion and Future Work

The field of biosurveillance is an ever-expanding science exploring methods to limit catastrophic disease outbreaks. Methods discussed in this paper range from disease diagnostic systems to forecasting cases using traditional time series methods, as well as new techniques adopted from statistical quality control. Techniques for evaluating these methods examine the timeliness of detection along with trade-offs between sensitivity and specificity of tests. For more information on the specific examination of these trade-offs, one should reference literature on receiver operating characteristic curves (ROC's). In recent years, systems have began using these same forecasting methods on data such as absenteeism, nurse hotline calls, and over-the-counter drug sales to detect disease outbreaks even sooner. Currently, the major obstacle to forecasting these data is defining some background behavior when no disease is present. Given time, forecasting these syndromic data will most certainly decrease detection time, saving lives and resources in the process.

References

[1] Aryel, R.M, Moore, A.W., Wagner, M.M., *Handbook of Biosurveillance.* Elsevier Academic Press. 2006

[2] Burkom, H.S., Murphy S.P., Shmueli, G., "Automated Time Series Forecasting for Biosurveillance" *Statistics in Medicine*

[3] Burkom, Howard & Shmueli, Galit "Statistical Challenges Facing Early Outbreak Detection in Biosurveillance." *American Statistical Association and the American Society for Quality.* February 2010, VOL. 52, NO.1

[4] Does, R.J.M.M., Meulen F.H., Vermat, M.B.,"Asymptotic Behavior of the Variance of the EWMA Statistic for Autoregressive Processes" *Science Direct* 3 January 2008

[5] Hohle, M. (2007). surveillance: An R package for the monitoring of infectious diseases. Computational Statistics, 22(4)

[6] Koshti, V.V. "Cumulative Sum Control Chart" *International Journal of Physics and Mathematical Sciences.* October-December 2011, VOL 1. NO. 1

[7] Mandi, Kenneth & Paola, Sebastiani "Biosurveillance and Outbreak Detection" Department of Biostatistics, Boston University

[8] Montgomery, D.C. *Introduction to Statistical Quality Control* New York: John Wiley and Sons. 2001

[9] Shephard, R. W., Toribio, B.A., Cameron, A.R., Thomson, P.C., Baldock, F.C.,"Development of the Bovine Syndromic Surveillance System (BOSSS)" Proceedings of the 11th International Symposium on Veterinary Epidemiology and Economics, 2006

[10] Zhang, N.F. "A statistical control chart for stationary process data" *Technometrics* 40, 24-38. 1998