Chapter 4

DATA ANALYSIS AND OUTBREAK DETECTION

The analysis components of a syndromic surveillance system focus on detecting the changes in public health status, which may be indicative of disease outbreaks. At the core of these analysis components is the automated process of detecting aberration or data anomalies in the public health surveillance data, which often have prominent temporal and spatial data elements, by statistical analysis or data mining techniques. These methods are also capable of dealing with various common problems in epidemiological data such as bias, delay, lack of accuracy, and seasonality. These techniques are the focus of this chapter.

When processing public health surveillance data streams, it is often necessary to map the collected syndromic data into a small set of syndrome categories to facilitate follow-up analysis and outbreak detection. Section 4.1 discusses related syndrome classification approaches. In Section 4.2, we provide a taxonomy of anomaly analysis and outbreak detection methods used for biosurveillance. Sections 4.3–4.6 summarize various specific detection methods spanning from classic statistical methods to data mining approaches, which quantify the possibility of an outbreak conditioned on surveillance data.

1. SYNDROME CLASSIFICATION

The onset of a number of syndromes can indicate certain diseases threatening public health. For example, the influenza-like syndrome could be due to an anthrax attack, which is of particular interest to biodefense. Syndrome

H. Chen et al., *Infectious Disease Informatics: Syndromic Surveillance for Public Health and BioDefense*, Integrated Series in Information Systems 21, DOI 10.1007/978-1-4419-1278-7_4,

classification thus is one of the first and important steps in syndromic data processing and analysis.

A substantial amount of research effort has been expended to classifying free-text chief complaints into syndromes. This classification task is difficult because different expressions, acronyms, abbreviations, and truncations are often found in free-text chief complaints (Sniegoski, 2004). For example, "chst pn," "CP," "c/p," "chest pai," "chert pain," "chest/abd pain," and "chest discomfort" can all mean "chest pain." On the basis of our summary findings reported in Section 3.1, a majority of syndromic surveillance systems use chief complaints as a major source of data. Therefore, the problem of mapping each chief complaint record to a syndrome category, referred to as syndrome classification, is an important practical challenge needing a solution. Another syndromic data type often used for syndromic surveillance purposes, i.e., ICD-9 or ICD-9-CM codes, also needs to be grouped into syndrome categories. Processing such information is somewhat easier as the data records are structured.

A syndrome category is defined as a set of symptoms, which is an indicator of some specific diseases. For example, a short-phrase chief complaint "coughing with high fever" can be classified as the "upper respiratory" syndrome. Table 4-1 summarizes some of the most commonly-monitored syndrome categories. Note that different syndromic surveillance systems may monitor different categories. For example, in the RODS system there are seven syndrome groups of interest for biosurveillance purposes, whereas EARS defines a more detailed list of 43 syndromes. Some syndromes are of common interest across different systems, such as respiratory or gastrointestinal syndromes.

Table 4-1. Diseases and syndrome categories commonly monitored.

Influenza-like	**Respiratory**	**Dermatological**
Fever	Neurologic	Cold
Gastrointestinal	Rash	Diarrhea
Hemorrhagic illness	Severe illness and death	Asthma
Localized cutaneous lesion	Specific infection	Vomit
Lymphadenitis	Sepsis	Other/none of the above
Constitutional		
Bioterrorism agent-related diseases		
Anthrax	Botulism-like/botulism	Plague
Tularemia	Smallpox	SARS (severe acute respiratory syndrome)

1.1 Syndrome Classification Approaches

The syndrome classification process can be either manual or implemented through an automated system. The BioSense system, developed by CDC (Ma et al., 2005), for instance, relies on a working group that develops syndrome mapping using CDC definitions. However, automated, computerized syndrome classification is essential to real-time syndromic surveillance. A software application that analyzes chief complaint records or ICD-9 codes and then determines appropriate syndrome categories is often known as a syndrome classifier.

Manual Grouping The BioSense system (Bradley et al., 2005; Sokolow et al., 2005) and the Syndromal Surveillance Tally Sheet program used in EDs of Santa Clara County, California, use a manual approach to classify the symptoms. They ask the medical experts in syndromic surveillance, infectious diseases, and medical informatics to perform the mapping of laboratory test orders into 11 syndromes categories defined by a multi-agency working group (Ma et al., 2005).

Automated Classification Existing automated classification methods can be roughly categorized into three groups: supervised learning, rule-based classification, and ontology-enhanced classification. The supervised learning methods require as input a set of CC records labeled with syndromes as learning samples before they can proceed to classify unlabelled CC records by syndromes. Naive Bayesian and Bayesian network-based methods are two examples of the supervised learning methods (Ivanov et al., 2002; Sniegoski, 2004). For instance, the CoCo chief complaints classifier developed as part of the RODS system is a Bayesian classifier (Chapman et al., 2003). Often, a learning approach has a natural language processing (NLP) component, which classifies free-text CCs with simplified grammar containing rules for nouns, adjectives, prepositional phrases, and conjunctions. As part of RODS, Chapman et al. adapted the MPLUS, a Bayesian network-based NLP system, to classify the free-text chief complaints (Wagner et al., 2004a; Chapman et al., 2005). Implementing learning algorithms is straightforward; however, collecting training records is usually costly and time-consuming. Another major disadvantage of supervised learning methods is the lack of flexibility and generalizability. Recoding for different syndromic definitions or implementing the CC classification system in an environment that is different from the one where the original labeled training data were collected could be costly and difficult.

In contrast, rule-based classification does not require labeled training data. A text string searching process for syndrome category classification is a typical rule-based approach. In general, the CC records are first cleansed and then mapped to the syndrome categories according to a set of rules often predefined by medical experts following the definitions of syndromes of interest. For instance, an example rule could be "fever, if NOT *animal* and NOT *environmental* and *fever*." Many applications, for example, EARS (Hutwagner et al., 2003), ESSENCE (CDC, 2003), and the National Bioterrorism Syndromic Surveillance Demonstration Program (Yih, Abrams et al., 2005), make use of such rules. Rule-based methods are relatively flexible, as the inference rules can be easily modified and updated. A major problem with rule-based classification methods is that they cannot handle symptoms not covered in the set of predefined rules.

The third category of automated approaches, ontology-based classification, utilizes relations between medical concepts (Leroy and Chen, 2001). Two representative methods are the BioPortal CC Classifier, which relies on Unified Medical Language System (UMLS) vocabularies and semantics (Lu et al., 2006, 2008), and the BioStorm approach, which uses a vocabulary abstraction method (Crubézy et al., 2005). BioPortal CC Classifier uses UMLS's Meta-thesaurus and SPECIALIST Lexicon to suggest a symptom grouping (as an intermediary representation) for a given CC record and then classify it using rules. It is able to provide a flexible architecture that supports easy adaptation to new syndromic categories. The BioStorm approach creates a series of intermediate abstractions up to a syndrome category from the primitive data (e.g., signs, lab tests) for syndromes indicative of illness due to an agent of bioterrorism.

We summarize representative syndrome classification methods in Table 4-2.

Table 4-2. Representative syndrome classification approaches.

Category	Example approaches	Application
Manual grouping	Medical experts perform the mapping of laboratory test orders into syndrome categories (Ma et al., 2005).	The BioSense system (Bradley et al., 2005; Sokolow et al., 2005) and Syndromal Surveillance Tally Sheet program in EDs of Santa Clara County, California.
Natural language processing (NLP)	NLP-based approaches classify free-text CCs with simplified grammar containing rules for nouns, adjectives, prepositional phrases, and conjunctions. Critiques of NLP-based methods include lack of semantic markings in chief complaints and the amount of training needed.	As part of RODS, Chapman et al. adapted the MPLUS, a Bayesian network-based NLP system, to classify the free-text chief complaints (Chapman et al., 2005; Wagner et al., 2004a).
Bayesian classifiers	Bayesian classifiers, including naïve Bayesian classifiers, bigram Bayes, and their variations, can classify CCs learned from the training data consisting of labeled CCs.	The CoCo Bayesian classifier from the RODS project (Chapman et al., 2003)
Text string searching	A rule-based method that first uses keyword matching and synonym lists to standardize CCs. Predefined rules are then used to classify CCs or ICD-9 codes into syndrome categories.	EARS (Hutwagner et al., 2003), ESSENCE (CDC, 2003), and the National Bioterrorism Syndromic Surveillance Demonstration Program (Yih et al., 2005)
Vocabulary abstraction	This approach creates a series of intermediate abstractions up to a syndrome category from the individual data (e.g., signs) for syndromes due to an agent of bioterrorism.	The BioStorm system (Crubézy et al., 2005; Buckeridge et al., 2002; Shahar and Musen, 1996)
Ontology-based classification	A rule-based system that can generalize symptoms grouping rules based on UMLS-derived vocabularies and semantics. It provides a flexible architecture for changing or adapting new syndromic categories.	The syndromic mapping component of the BioPortal system (Lu et al., 2008)

An interesting complementary method using both manual and natural-language processing techniques to create CC classifiers is presented by Halasz et al. (2006). They apply an *n*-gram text processing program to build an ICD9 classifier to a training set of ED visits for which both the CC and ICD9 code are known. A collection of CC substrings with associated probabilities was constructed and used to generate a CC classifier program. This approach allows the rapid automated creation and updating of CC classifiers based on ICD9 groupings.

Researchers have also started working on a CC classifier for non-English CCs. It is noted that there is a critical need for the development CC classification systems capable of processing non-English CCs as syndromic surveillance is being increasingly practiced around the world. One design first maps non-English CCs to English CCs and then use well-tested English CC classification systems to process translated CCs (Lu et al., 2007a).

1.2 Performance of Syndrome Classification Approaches

On the basis of our survey, about 40% of syndromic surveillance systems use automated syndrome classification, while the other 40% rely on a manual approach (details are unknown for the remaining 20%). There is clearly room for improvement and adoption of automated methods.

Evaluation studies have been conducted to compare various classifiers' performance for selected syndrome types (Travers and Haas, 2004). For instance, experiments comparing two Bayesian classifiers for the acute gastrointestinal syndrome showed a 68% mapping success against expert classification of ED reports (Ivanov et al., 2002). In general, however, it is difficult to paint a general picture of how well syndromic classifiers perform and how they fare against each other as many systems have not been evaluated on classification accuracy. In addition, the performance of these classifiers varies with different syndrome categories, further complicating the evaluation task.

Many prior studies show that a considerable portion (30–40%) of the chief complaints data is not classifiable because they are too noisy. However, combining chief complaints with the diagnostic codes (such as ICD-9) during the same visit can achieve a better classification accuracy (Reis and Mandl, 2004).

Another challenge facing syndrome classification is that there are no universally-accepted, standardized syndrome definitions. As a result, significant rewriting/fine-tuning efforts are needed when applying a classification approach in particular application contexts. One possible approach to deal with these difficulties is to create intermediary representations (such as symptom groups)

and create explicit rules that map these intermediary representations into customized syndrome categories (Lu et al., 2006).

2. A TAXONOMY OF OUTBREAK DETECTION METHODS

Syndromic surveillance systems typically make available multiple outbreak detection algorithms, as no single method can deliver superior performance across a wide range of scenarios or meet different surveillance objectives (Buckeridge et al., 2003).

Many statistical and data mining techniques for syndromic surveillance have been proposed in the literature. These methods can be generally divided into retrospective and prospective approaches. If instead we consider the characteristics of the surveillance data analyzed, another orthogonal classification scheme is possible, dividing the outbreak detection methods into temporal analysis, spatial analysis, and spatial-temporal analysis approaches. This subsection focuses on both schemes.

Interested readers are referred to http://statpages.org/, which provides tutorials for various kinds of parametric and nonparametric statistical tests that form the statistical foundation of outbreak detection, and http://www.autonlab.org/tutorials/, which includes statistical data mining and machine learning tutorials. The review articles on data mining and its application in health and medical information (Bath, 2004; Benoit, 2002) are also good references to provide in-depth background for the material presented in this section.

2.1 Retrospective vs. Prospective Syndromic Surveillance

A number of surveillance approaches fall under the general umbrella of *retrospective* models, which aim at testing statistically whether events are randomly distributed over space and time for a predefined geographical region during a predetermined time period (Kulldorff, 2001). Some examples of retrospective methods include space scan statistic (Kulldorff, 1997), Nearest Neighbor Hierarchical Clustering (NNH) (Levine, 2002), and Risk-adjusted Support Vector Clustering (RSVC) (Zeng et al., 2004a). When applying retrospective methods, there is usually a clear distinction between the baseline data points and the observations of interest, where the baseline data correspond to known "normal" health status and the observations of interest are case reports to be examined for surveillance purposes. In applications where the separation between the baseline data and observations of interest can be

cleanly and meaningfully done, retrospective methods can be effectively applied.

One major limitation of retrospective methods is that they are slow in detecting emerging clusters when the separation between the baseline data and observations of interest is not obvious. The resulting manual trial-and-error interventions severely limit the applicability of retrospective methods.

Prospective surveillance often entails repeated analyses performed periodically on incoming surveillance data streams to identify statistically significant changes in an online context (Chang et al., 2005). Using such a method, the separation of the baseline data and observations of interest is no longer needed as the system automatically tries various combinations of having some time windows as the baseline and some periods after them as the time of interest.

Prospective analysis has long been used in disease surveillance applications. The CUSUM method is one of the most established methods. Other examples include Rogerson's approaches (Rogerson, 1997), Kulldorff's prospective version of time-space scan statistics (Kulldorff, 2001), and the Prospective Support Vector Clustering (PSVC) method (Chang et al., 2005).

2.2 Temporal, Spatial, and Spatial-Temporal Outbreak Detection Methods

Table 4-3 summarizes a wide range of outbreak detection methods, all of them implemented in one or more syndromic surveillance systems surveyed. They are divided into three groups: temporal, spatial, and spatial-temporal (Buckeridge et al., 2005b; Mandl et al., 2004). Note that this table does not attempt to exhaustively list every detection algorithm proposed in the literature. Interested readers can refer to (Brookmeyer and Stroup, 2004; Lawson and Kleinman, 2005) for recent in-depth reviews of a more comprehensive set of algorithms. The methods listed in Table 4-3 are chosen because of their connection with the syndromic surveillance systems surveyed. Although not exhaustive, it covers most of the detection method types and provides a useful snapshot of the state of the art. Sections 3-5 provide additional analysis of these three groups of detection methods, respectively.

Table 4-3. Outbreak detection algorithms.

Algorithm	Short description	Availability and applications	Features and problems
Temporal analysis			
Serfling method	A static cyclic regression model with predefined parameters optimized through the training data	Available from RODS (Tsui et al., 2001); used by CDC for flu detection; Costagliola et al. applied Serfling's method to the French influenza-like illness surveillance (Costagliola et al., 1981)	The model fits data poorly during epidemic periods. To use this method, the epidemic period has to be predefined.
Autoregressive Integrated Moving Average (ARIMA)	A linear function learns parameters from historical data. Seasonal effect can be adjusted.	Available from RODS	Suitable for stationary environments.
Recursive Least Square (RLS)	A dynamic autoregressive linear model that predicts the current count of each syndrome within a region based on the historical data; it continuously adjusts model coefficients based on prediction errors	Available from RODS	Suitable for dynamic environments.
Exponentially Weighted Moving Average (EWMA)	Predictions based on exponential smoothing of previous several weeks of data with recent days having the highest weight (Neubauer, 1997)	Available from ESSENCE	Allowing the adjustment of shift sensitivity by applying different weighting factors.

Algorithm	Short description	Availability and applications	Features and problems
Cumulative Sums (CUSUM)	A control chart-based method to monitor for the departure of the mean of the observations from the estimated mean (Das et al. 2003; Grigoryan et al., 2005). It allows for limited baseline data.	Widely used in current surveillance systems including BioSense, EARS (Hutwagner et al., 2003) and ESSENCE, among others	This method performs well for quick detection of subtle changes in the mean (Rogerson 2005); it is criticized for its lack of adjustability for seasonal or day-of-week effects.
Hidden Markov Models (HMM)	HMM-based methods use a hidden state to capture the presence or absence of an epidemic of a particular disease and learn probabilistic models of observations conditioned on the epidemic status.	Discussed in (Rath et al., 2003)	A flexible model that can adapt automatically to trends, seasonality covariates (e.g., gender and age), and different distributions (normal, Poisson, etc.).
Wavelet algorithms	Local frequency-based data analysis methods; they can automatically adjust to weekly, monthly, and seasonal data fluctuations.	Used in NRDM to indicate zip-code areas in which OTC medication sales are substantially increased (Espino and Wagner 2001; Zhang et al., 2003)	Account for both long-term (e.g., seasonal effects) and short-term trends (e.g., day-of-week effects) (Wagner et al., 2004b).
Spatial analysis			
Generalized Linear Mixed Modeling (GLMM)	Evaluating whether observed counts in relatively small areas are larger than expected on the basis of the history of naturally occurring diseases (Kleinman et al., 2005a; Kleinman et al., 2004)	Used in Minnesota (Yih et al., 2005)	Sensitive to a small number of spatially focused cases; poor in detecting elevated counts over contiguous areas when compared with scan statistic and spatial CUSUM approaches (Kleinman et al., 2004).

SMall Area Regression and Testing (SMART)	An adaptation of GLMM that takes into account multiple comparisons and includes parameters for ZIP code, day of the week, holiday, and seasonal cyclic variation.	Available from BioSense and National Bioterrorism Syndromic Surveillance Demonstration Program (Yih et al., 2005)	Seasonal, weekly effects, and other parameters under consideration can be adjusted during the regression process.
Spatial scan statistics and variations	The basic model relies on using simply-shaped areas to scan the entire region of interest based on well-defined likelihood ratios. Its variation takes into account factors such as people mobility	Widely adopted by many syndromic surveillance systems; a variation proposed in (Duczmal and Buckeridge 2005); visualization available from BioPortal (Zeng et al., 2004a).	Well-tested for various outbreak scenarios with positive results; the geometric shape of the hotspots identified is limited.
Bayesian spatial scan statistics	Combining Bayesian modeling techniques with the spatial scan statistics method; outputting the posterior probability that an outbreak has occurred, and the distribution of this probability over possible outbreak regions	Available from RODS (Neill et al., 2005)	Computationally efficient; can easily incorporate prior knowledge such as the size and shape of outbreak or the impact on the disease infection rate.
Spatial-temporal analysis			
Space-time scan statistic	An extension of the space scan statistic that searches all the sub-regions for likely clusters in space and time with multiple likelihood ratio testing (Kulldorff 2001).	Widely used in many community surveillance systems including the National Bioterrorism Syndromic Surveillance Demonstration Program (Yih et al., 2004)	Regions identified may be too large in coverage.

Algorithm	Short description	Availability and applications	Features and problems
What is Strange About Recent Event (WSARE)	Searching for groups with specific characteristics (e.g., a recent pattern of place, age, and diagnosis associated with illness that is anomalous when compared with historic patterns) (Kaufman et al. 2005)	Available from RODS; Implemented in ESSENCE	In contrast to traditional approaches, this method allows for use of representative features for monitoring (Wong et al. 2003; Wong et al. 2002). To use it, however, the baseline distribution has to be known.
Population-wide ANomaly Detection and Assessment (PANDA)	A causal Bayesian network approach to model a population and infer the spatial-temporal probability distribution of disease for the entire population or individual patients	Available from RODS (Cooper et al. 2004; Moore et al. 2002)	Extensive computational effort
Prospective Support Vector Clustering (PSVC)	This method uses the Support Vector Clustering method with risk adjustment as a hotspot clustering engine and a CUSUM-type design to keep track of incremental changes in spatial distribution patterns over time	Developed in BioPortal (Chang et al. 2005; Zeng et al. 2004a)	This method can identify hotspots with irregular shapes in an online context

Because of the importance of outbreak detection algorithms for syndromic surveillance, we review some of the critical methods adopted in more detail below. The readers should note that the models we are about to discuss can be written in a number of mathematically equivalent ways, while the ones presented in the text are one of the representations.

3. TEMPORAL DATA ANALYSIS

This section discusses representative temporal anomaly detection methods. Temporal anomaly detection belongs to the vast domain of time series analysis. It monitors public health events or incidences as a sequence of data points, measured typically at evenly-distributed successive times. Temporal anomaly detection methods attempt to identify unusual patterns, smooth out naturally-occurring (or known) variations, and distinguish the variations caused by a possible outbreak from natural variations. Such methods either study the event frequency or the intensity of adverse event occurrences (time intervals between occurrences) to detect changes. These changes could follow different trends (e.g., linear, exponential).

3.1 Statistical Process Control (SPC)-Based Anomaly Detection

A majority of the systems surveyed employ statistical process control (SPC)-based algorithms. These algorithms were originally developed to monitor a process and its mean in industrial settings. The ability to differentiate the "out-of-control" mean from the "in-control" mean makes these methods readily applicable for anomaly detection.

The basic idea behind SPC-based algorithms is as follows. A small random sample $x = (x_1, ...x_t, ...)$ is drawn repeatedly at certain time intervals. The sample mean is compared against given thresholds; alarms are triggered at $t_A = \min\{s; \text{sample\_mean}(x_s) > G(s)\}$, if the sample mean exceeds the control limit *G(s)*. The alerting threshold is either theoretically defined, or dynamically estimated through historical data. The later one is proved to be more robust than the former (Buckeridge et al., 2005a). The single time-series analyzed often exhibits substantial day-of-week or seasonal patterns. As such, it is a common practice to estimate the incidence rate using a linear or Poisson regression model, and then to apply a SPC-based method to the regression residuals (Buckeridge et al., 2005a).

The Control Statistical Cumulative Sums (CUSUM) and Exponentially Weighted Moving Average (EWMA) methods are two standard SPC-based methods that have been widely applied for outbreak detection. CUSUM

keeps track of the accumulated deviation between observed and expected values. Formally, the accumulated deviation is defined as $S_t = \max(0, S_{t-1} + z_t - k)$, where k is a control parameter and z_t models the distribution of the variable of interest (e.g., $z_t = \frac{x_t - \mu_t}{\sigma_t}$, if the variable is normally distributed) (Rogerson, 2005). Different forms of CUSUM have been developed, which assume that the underlying distribution could be Poisson or exponential (Rogerson, 2005). Nonparametric models have also been developed, removing the need for knowledge of the underlying distribution. A deployed SPC method often incorporates a short guard band (e.g., 2 days) between the baseline period and the day to be monitored. The guard band may lift the sensitivity by avoiding a gradually increasing outbreak contaminating the baseline with the outbreak signal. CUSUM methods have been specifically designed to deal with limited availability of historical data. Three CUSUM algorithms used in the EARS system require less than 10 days as the baseline period. They differ from each other by the different settings of the baseline period and the threshold levels, resulting in different levels of sensitivity (Hutwagner et al., 2003).

The Shewhart method is another simple form of SPC-based methods. It can be viewed as performing repeated significance tests on deviations of an observation from a target constant. The Shewhart method performs poorly for small and moderate shifts, but for large shifts, CUSUM actually converges to the Shewhart method (Lawson and Kleinman, 2005). One study used a Shewhart control chart to detect epidemics of Influenza A (Quenel et al., 1994).

Instead of considering only the last observation in the Shewhart method, the exponentially weighted moving average (EWMA) method monitors all the previous observations, summing up the multiple deviations in a weighted scheme, giving the most recent observation the greatest weight, and all the previous observations geometrically decreasing weights (Neubauer, 1997).

SPC-based methods are widely used in surveillance due to their simplicity. Their performances have been tested in many real settings. BioSense, EARS, and ESSENCE syndromic surveillance systems among others implemented either CUSUM or EWMA or both, and reported their early aberration detection capacity for influenza-like illness and other diseases (Hutwagner et al., 2005a; Zhu et al., 2005). The details of the performance evaluation can be found in Chapter 6.

3.2 Serfling Statistic

Serfling's method uses cyclic regression to model the normal pattern of the numbers of patients susceptible to death for pneumonia and influenza when there is not an epidemic with the objective of determining an epidemic

threshold. Its use requires a clear definition of the disease, the selection of data to identify a normal pattern of susceptible patients, and the assumption that the normal pattern is periodical.

The Serfling statistic was originally proposed by Serfling for statistical analysis of weekly pneumonia and influenza deaths in 108 US cities in 1963 (Serfling, 1963). Serfling's method uses cyclic regression to establish an expected threshold for daily statistic based on history data excluding the epidemic weeks, accounting for seasonal variations. It requires a clear definition of the disease and the assumption that the normal pattern is periodical (Mandl et al., 2004). A theoretical form of this method is formulated as:

$$y(t) = c_1 + c_2 t + c_3 \sin(2\pi \frac{t}{52}) + c_4 \cos(2\pi \frac{t}{52})$$

Serfling's method is regarded as a traditional modeling technique applied to a number of disease surveillance practices such as the French influenza-like syndrome data (Costagliola et al., 1981). Serfling's method has also been used by RODS system to model hospital visitation data for influenza (Tsui et al., 2003).

3.3 Autoregressive Model-Based Anomaly Detection

The autoregressive integrated moving average (ARIMA) method is a class of time-series analysis models that are typically specified by three parameters: the order of autocorrelation (AR), the order of integration (I), and the order of moving average (MA) (Box et al., 1994). These parameters determine two things: how much of the past should be used to predict the next observation and how much do the past observations weigh in predicting the next observation. The higher-order models are more complex and can usually achieve a better fit of the training data set, while the simpler low-order models are usually less likely to over-fit to training dataset (Reis and Mandl, 2003). Description of the class of ARIMA methods in full details can be found in (Box et al., 1994). We here give an example ARIMA (1, 1, 1) model to simply show the notations. In the following equation, μ is a constant term, $(Y(t-1) - Y(t-2))$ represents a first-order "autoregressive" term, and the forecast error – first-order moving average at period $t-1$ is $e(t-1)$. ϕ and θ are coefficients.

$$\hat{Y}(t) = \mu + Y(t-1) + \phi(Y(t-1) - Y(t-2)) - \theta e(t-1)$$

ARIMA models have been applied to pneumonia and influenza deaths for detection of outbreaks (Reis and Mandl, 2003). In the Automated Epidemiologic Geotemporal Integrated Surveillance (AEGIS) program at Children's Hospital Boston and Harvard Medical School, a hybrid of ARIMA with cyclic regression was found to have excellent predictive ability (Mandl et al., 2004). These models are available in many common statistical software packages (e.g., SAS Time Series Forecasting module). One drawback of the ARIMA models is that there is no systematic way to update model parameters when new data points arrive.

The Recursive Least Square (RLS) algorithm is another method based on autoregressive linear models and is implemented as part of RODS (Wong et al., 2002, 2003). It learns from the time series but does not need a large learning sample. Also it is more sensitive to recent historical data to predict outcomes, so it is well suited to surveillance for short-term events. Unlike ARIMA or the Serfling method, RLS continuously updates its parameters. RLS operates by converging on a set of coefficients (for a weighted linear equation) that best predicts historical values. The algorithm uses these coefficients to predict the current value. It calculates the prediction errors between the predicted values and the time series values. Using the prediction errors and algorithm threshold (expressed in number of standard deviations), RLS computes a threshold value. This algorithm is ideal for detecting spikes of cases when there is little historical data. Using these models implies that transformation of the data leads to a stationary time series, for which a single underlying probability distribution is assumed. These two hypotheses are not necessarily true, however; the data may present abrupt and wide changes of magnitude as well as irregular periodicity, in situations such as epidemics, modifications of the case-definition, screening, or vaccination (Le and Carrat, 1999).

3.4 Hidden Markov Model (HMM)-Based Models

The SPC-based models and the cyclic regression methods need nonepidemic data to model the baseline distribution, which is not always available without data preprocessing. This makes it an obstacle for automated surveillance. Researchers, therefore, have proposed to use Hidden Markov Models (HMM) to segment the time series of influenza indicators into epidemic and nonepidemic phases. Hidden Markov models have found major success in temporal pattern recognition such as speech and handwriting recognition, and bioinformatics. The basic idea behind HMM-based models is to add

another layer of random signal generation process conditioned on the state of a hidden Markov process to determine the conditional distribution of each observed data point.

The sequence of state transitions in HMM is reconstructed using statistical methods to calculate the most likely trends of the surveillance data. HMM-based models are flexible enough to be easily adapted automatically to trends, seasonality, covariates (e.g., gender and age), and different distributions (normal, Poisson, Gaussian, Gamma, etc.). HMM-based models have been applied in a number of surveillance data time series analysis studies. For example, Le Strat and Carrat applied a univariate HMM to ILI time series surveillance in France (Le and Carrat, 1999). More technical details of HMM in disease surveillance can be found in (Madign, 2005). The author further discussed the proper number of hidden states, multivariate extensions to the above univariate HMM, as well as HMMs with random observation times. Madigan also pointed out that a key extension to the existing research on HMM-based surveillance would be to incorporate a spatial component in the hidden layer of the models.

4. SPATIAL DATA ANALYSIS

Spatial analysis techniques are used to find the extent of "clustering" of cases across a map and have long been an important component of the surveillance analysis toolset. More specifically, spatial clustering analysis aims to detect and locate the anomalies in disease occurrences or outbreaks by examining the surveillance data's spatial distribution, as clusters might be of insufficient size to be detected in analyses that consider only an entire region. This would also allow for the possibility that some areas contained populations more likely to become sick, such as older people, or more likely to seek healthcare, as might be the case for certain cultural groups. It thus provides the capability of tracking the progression of disease outbreaks and identifying the population at risk for proper treatment and prevention.

The rationale behind spatial surveillance is that natural disease outbreaks or biological attacks are typically localized at some spatial scale. Spatial analysis in syndromic surveillance uses spatial information residing in the data, such as the patient's home residence, sometimes the work place, and the location of the hospital where the illness is reported. Temporal analyses we discussed in the earlier section are capable of detecting elevated rates across an entire region, but would be less sensitive to a smaller number of spatially focused cases. Furthermore, spatially correlated random effects are

often ignored by pure time series methods, thus it is assumed that all tests are independent.

Investigations of clusters in space often associate the varying population density with the null hypothesis. Denote the intensity of the disease cases (the number of expected events per unit area) by $\lambda_0(s)$, where s represents a location in the study area. Also denote by $\lambda_1(s)$ the intensity function of the population at risk. The null hypothesis of normal spatial distribution is in fact a proportional intensity function, $H_0 : \lambda_0(s) = \rho\lambda_1(s)$, where ρ is the expected number of cases divided by the expected number at risk.

One widely-used spatial analysis algorithm is SMART, made available through the BioSense system and the National Bioterrorism Syndromic Surveillance Demonstration Program. Other popular methods include the GLMM algorithm (Kleinman et al., 2004); spatial scan statistics (Kulldorff, 1999) and a number of its variations such as Modified spatial scan statistics (Duczmal and Buckeridge, 2005); and the Risk-adjusted Support Vector Clustering (RSVC) method (Zeng et al., 2004a).

Temporal analysis methods such as CUSUM can also be adapted to analyze spatial information by maintaining CUSUM charts for the surrounding neighborhood of each individual region as local spatial statistics or by maintaining multivariate CUSUM charts for all regions in a global setting (Lawson and Kleinman, 2005). Vice versa, spatial clustering techniques could be adapted to temporal surveillance, if considering time as one-dimensional space.

4.1 Generalized Linear Mixed Models and SMART Algorithm

Kleinman et al. (2005a) proposed the use of Generalized Linear Mixed Model (GLMM) statistics based on a logistic regression model to estimate the probability that each subject under surveillance is a case, in each area, on a given day. The simple logistic regression model introduces "shrinkage" estimators showing the density of population in each area, as the size of the population under surveillance in each area often varies. The proposed method treats each small area as if it was an individual, and the relative locations of the small areas are not taken into account by the model. This method in essence ignores much spatial information and cannot detect elevated counts over several contiguous areas.

SMART is an adaptation of the GLMM method, taking additional parameters into account to adjust for seasonal, weekly, social trends, and holiday status (Bradley et al., 2005). In such an approach, generalized linear models are used to establish the expected count per ZIP code per day based on regressing historical series of counts in each small area. The established

distribution of case counts are then refined to account for multiple ZIP codes through multiple testing. One experimental study suggested that SMART delivered slightly inferior results to the spatial scan statistic method. However, both methods achieved good performances (Kleinman et al., 2005a).

4.2 Spatial Scan Statistic and Its Variations

Most syndromic surveillance systems make use of spatial scan statistic and its variations. Using such methods for spatial analysis, a large set of circular windows with varying sizes is imposed on the map in different locations to search for clusters over the entire region. As the cluster size is unknown a priori, the scan statistic method uses a likelihood ratio test where the alternative hypothesis is that there is an elevated rate within the scanning window when compared with outside. The most likely clusters can then be identified based on the likelihood-ratio test if the null hypothesis is rejected. For each distinct window, the likelihood ratio is proportional to: $(\frac{n}{\mu})^n (\frac{N-n}{N-\mu})^{N-n}$, where n is the number of cases inside the circle, N is the total number of cases, and μ is the expected number of cases inside the circle (Kulldorff, 1997). Other probability models, i.e., distribution from which the case incidence are generated, have also been used for scan statistics. Poisson model is commonly seen. Bernoulli model can be used for on-off case-control type data, and exponential model is for survival data.

There are several advantages with scan statistic methods. First, they avoid preselection bias regarding the size or location of clusters. Second, they can be easily adjusted for nonuniform population density as well as other factors such as age.

The spatial-temporal version of the scan statistic uses cylinders instead of circles, where the height of the cylinder represents time. Still, the circular base defines a geographic area with a varying radius. The size of the area that is circled could be from zero to hundreds of kilometers or everything in between. The height of the cylinder can represent a time of day or years. The rest of the process is largely unchanged. A moving cylindrical window with variable sizes in both space and time visits all spatial-temporal locations to identify a significant excess of cases within it, until it reaches a predetermined size limit (Kulldorff, 1999, 2001). On the basis of the flexible purely spatial scan statistic, Takahashi et al. proposed a flexibly shaped space-time scan statistic for detecting irregularly-shaped clusters, which may not be detected by the circular spatial scan statistic (Takahashi et al., 2008). The performance of the flexibly-shaped space-time scan statistic is compared with the cylindrical scan statistic with a space-time power distribution

developed by extending the purely spatial bivariate power distribution (Takahashi et al., 2008).

SaTScan is a freely-available software package that implements various types of spatial and space-time scan statistics (2006j). It has been used in more than 10 syndromic surveillance systems, according to our survey. Two commercial products, WpiAnalyst extension for ArcView GIS from the Public Health Research Laboratories (2003d) and ClusterSeer developed by TerraSeer (2006c) contain both spatial and spatial-temporal scan statistics together with many other statistical clustering methods. The SaTScan Macro Accessory for Cartography (SMAC) package consists of four SAS macros and was designed as an easier way to run SaTScan multiple times and add graphical output. The package contains individual macros, which allow the user to make the necessary input files for SaTScan, run SaTScan, and create graphical output all from within SAS software. The macros can also be combined to do this all in one step (Abrams and Kleinman, 2007).

A modified spatial scan statistic proposed by Duczmal and Buckeridge considers work-related factors. A factor reflecting the number of "contaminations" from workers at the nearest neighbors is added to the observed cases in the residential zones (Duczmal and Buckeridge, 2005). Their simulation shows that their approach can achieve greater detection power than the scan statistics that do not consider people movements. To apply their approach, workplace location information is required, which unfortunately is not commonly available in surveillance data sources.

There are a few known problems with spatial scan methods. First, they can only identify clusters in simple regular shapes. Second, it is difficult to incorporate prior knowledge, such as the size or shape of the outbreaks or the impact on disease infection rate. Third, exhaustive searches over a large region to perform statistical tests could be computationally expensive.

The method summarized in the next subsection deals with the first problem. To address the second and third problems, Neill et al. (2005) proposed a Bayesian spatial scan statistic that is computationally more efficient and capable of combining the a priori knowledge of the investigated outbreak. A conjugate Gamma-Poisson model, as opposed to the Poisson model in Kulldorff's original spatial scan statistic, is used to produce a spatially smoothed map of disease rates, with a focus on computing the posterior probabilities to determine the outbreak likelihood and to estimate the location and size of potential outbreaks.

4.3 Risk-Adjusted Support Vector Clustering (RSVC) Algorithm

Zeng et al. developed an approach called RSVC that combined the risk adjustment idea with a robust Support Vector Clustering (SVC) method to improve the quality of retrospective spatial-temporal analysis. Specifically, for regions with prior dense baseline data distribution, data points are less likely to be grouped to form anomaly clusters. Several steps are involved in the clustering process. First, the input data are implicitly mapped to a high-dimensional feature space defined by a kernel function (typically the Gaussian kernel). Second, the algorithm finds a hypersphere in the feature space with a minimal radius to contain most of the data. The problem of finding this hypersphere can be formulated as a quadratic or linear programming problem depending on the distance function used. Third, the function estimating the support of the underlying data distribution is then constructed using the kernel function and the parameters learned in the second step. The width parameter in the Gaussian kernel function is dynamically adjusted based on kernel density computed using background data. When mapped back to original space, the hypersphere splits into several clusters, which indicated high risk outbreak areas (Zeng et al., 2004b).

5. SPATIAL-TEMPORAL DATA ANALYSIS

5.1 Rule-Based Anomaly Detection with Bayesian Network Modeling

The "What's Strange About Recent Events" (WSARE) algorithm performs a heuristic search over combinations of temporal and spatial features to detect irregularities in space and time. The case features analyzed by WSARE include syndrome category, age, gender, and geographical information. For example, a two-term case feature could be "Gender = Male AND Home Location = NW." The number of the cases satisfying and those not satisfying the case feature are computed to be used to determine whether there is significant discrepancy between the observed statistic of the current day and the baseline.

Historic data (e.g., recent weeks before the day of analysis) is fed to a Bayesian network to create a baseline distribution. The network is constructed using an algorithm called optimal reinsertion (Moore et al., 2003) based on ADTrees (Moore and Lee, 1998). The benefit of the approach relies on Bayesian network's generalization capability that is able to predict the probability of a situation that may not have been encountered in the past. The network structure is rebuilt every month, while the parameters are updated

daily. Environmental attributes such as season and day of week can be incorporated in the model as conditional probability.

All feature-value combinations are then searched and scored exhaustively. The scores are generated by conducting hypothesis testing for each feature-value combination against the baseline distribution. Instead of exhaustively searching for *i*-term feature-value combinations with an exponential complexity (i = 1, 2, …, n, suppose that there are n features in total), a greedy search approach is designed by searching the best 1-term case feature first and then adding another term to it to compose a 2-term case feature, and so forth. Compared with several other algorithms that do not examine covariate information, WSARE performed better as measured by timeliness at the expense of slightly higher false-positive rate (Wong et al., 2002).

5.2 Population-Wide Anomaly Detection and Assessment (PANDA)

Population-Wide Anomaly Detection And Assessment (PANDA) is a causal Bayesian network-based model constructing and inferring the spatial-temporal probability distribution of disease in a population as a whole. The causal Bayesian network consists of a large set of inter-linked patient-specific probabilistic causal models, each of them including variables that represent risk factors (e.g., infectious disease exposures of various types), disease states, and patient symptoms (Cooper et al., 2004). Simulation conducted by the RODS team showed that the model can handle a population size of 1.4 million (Cooper et al., 2004).

6. MONITORING MULTIPLE DATA STREAMS

In Sections 6 and 7, we discuss two specific sets of issues concerning outbreak detection that are worth separate treatments.

In disease surveillance, multiple data sets (data are collected simultaneously from pharmacies, hospitals, nurse help telephone calls, and clinics) are usually available for surveillance. However, the majority of implemented detection algorithms monitor individual data sources and do not cross reference between them. The problem is that no single data source captures all the individuals in the outbreak (Kulldorff et al., 2005). One potentially fruitful detection approach is a data-fusion approach using multiple sources of data (e.g., ED visits and OTC sales data) to perform outbreak detection. For example, MCUSUM and MEWMA (Yeh et al., 2003, 2004) were developed to increase detection sensitivity while limiting the number of false alarms. Multiple univariate statistical techniques and multivariate methods have also

been used in prior studies based on different independence assumptions among the data streams. Multiple univariate methods assume independence among the data; while multivariate methods establish the covariance matrix typically estimated from a baseline period (Buckeridge et al., 2005a). In the ESSENCE II project, chief complaints data and sales of OTC medications are treated as covariates (Lombardo et al., 2004). However, to model the multiple univariate signals from different data streams, an in-depth investigation and characterization of health-care-seeking behavior is necessary.

Another approach is to monitor stratified data (e.g., based on syndrome type or age group, counties, or treatment facilities) in parallel. The WSARE (What is Strange About Recent Events) system proposed by Wong et al. (2003) is one example, which searches for outbreaks in various groupings of age, gender, or census tracts. Kulldorff et al. (2003) developed a tree-based scan statistic to do surveillance on groupings that can be preclassified into a hierarchical tree structure.

In addition, during major public events, unpredictable shifts in the healthcare data may occur due to changes in healthcare utilization patterns. This problem is addressed by Reis et al. Instead of monitoring different healthcare data streams individually, they proposed a class of epidemiological network models that monitor the interrelationships among these data streams. The integrated network-based modeling of the interrelationships among the epidemiological data streams allows more robust performance in the face of shifts in healthcare utilization during epidemics and major public events (Reis et al., 2007).

Simultaneous wavelets analysis over multiple time series are practiced by Dillard and Shmueli (Shmueli and Fienberg, 2006). Rigorous comparative evaluations to quantify the gain of using covariates from multiple data sources in surveillance are needed.

7. SPECIAL EVENTS SURVEILLANCE

Another challenging issue for real-time outbreak detection is that the surveillance algorithms often rely on historic datasets that span a considerable length of time. Few methods demonstrate reliable detection capability with short-term baseline data. This is a particular concern for surveillance systems for special events (also referred to as drop-in models), which are implemented against bioterrorism attacks or natural disease outbreaks in settings such as international and national sports events or meetings that involve many participants in a short time window.

EARS was used for syndromic surveillance at several large public events in the United States, including the Democratic National Convention of 2000, the 2001 Super Bowl, and the 2001 World Series (Hutwagner et al., 2003).

The RODS system was used during the 2002 Winter Olympic Games (Gesteland et al., 2002). The LEADERS system often serves as a drop-in surveillance system intended to facilitate communication and coordination within and between public health facilities (Ritter, 2002).

8. SUMMARY OF DATA ANALYSIS PROCESS FOR SYNDROMIC SURVEILLANCE

In this chapter, we first introduce syndrome classification as the first step of syndromic data analysis. We then summarize a large number of disease surveillance algorithms. These algorithms are organized in two dimensions. In the first dimension, a surveillance method is either retrospective surveillance or prospective. Retrospective analysis focuses on analyzing historical data, whereas prospective analysis is more useful for processing online data streams. In the second dimension, a surveillance method can be seen as either a temporal, spatial, or spatial-temporal analysis method. Methods designed for special events are discussed separately due to their unique characteristics. We also examine methods that monitor multiple data streams, which warrant further exploration due to their importance and applicability. We conclude this chapter by pointing out some technical issues to watch for while applying these surveillance methods.

First, the outbreak detection methods make a number of assumptions about the analyzed data. The distribution of the disease events are in many cases assumed, so before the application of any surveillance methods to the disease data, there should be analysis regarding disease behaviors such as the outbreak patterns and events distribution. Second, an algorithm's performance is related to a number of settings: (1) the availability of historic data; data collection process as discussed in Chapter 2 is thus closely related to a surveillance algorithm performance; (2) the type of outbreak signals (e.g., slow-building or surge outbreak); (3) the spatial granularity of the data in spatial analysis.

All the complications due to the dynamics of different diseases need to be considered and well investigated before applying a detection algorithm. In (Burkom and Murphy, 2007), the authors propose a data-adaptive method selection scheme to "suit the remedy to the case," by first evaluating a number of data discriminates such as mean, variance, and skewness before selecting a detection algorithm for analysis. The BioStorm research group developed an ontology-based method to incorporate the a priori knowledge so that different analytical methods are assigned to different types of surveillance data in different settings (Crubézy et al., 2005).