Data Analysis Research Issues and Emerging Public Health Biosurveillance Directions

Henry Rolka

National Center for Public Health Informatics, Centers for Disease Control and Prevention, HRolka@cdc.gov

Statistical and data analysis issues for new surveillance paradigms are quickly emerging in public health. Among the key factors motivating their evolution and development are

1. New requirements and resources to address a perceived bioterrorism threat as well as emerging diseases.
2. Information system technology growth in general.
3. Recognition of surveillance integration as a priority.
4. Widely available data with unrealized potential for useful information.

The term *syndromic surveillance* is used here somewhat as a catch-all for referring to new surveillance system paradigms and should be interpreted broadly [MH03, Hen04]. Biosurveillance has a much longer history for naturally occurring morbidity and mortality (i.e., infectious diseases, birth defects, injuries, immunization coverage, sexually transmitted diseases, HIV, medical product adverse events, etc.) than for deliberately malicious exposures. The professional relationships and established roles among public health levels (local, state, and federal) must be considered carefully as the context in which public health surveillance activity and system maturity take place. To ignore this extant infrastructure in advancing surveillance methodology involves the risk of developing irrelevant ideas because they may not be feasible to implement. However, if we do not extend beyond our applied creativity, we risk stagnation and incompetence. The balance is to identify the right size research, development, and implementation steps that will enable palatable progress and then take these steps quickly, frequently, and repeatedly in the same direction. Implementation of national scope public health information system change is extremely complex. This is a prologue designed to introduce and sensitize the reader to factors that may serve as enablers in that complexity for taking advantage of how to best consider the concepts asserted in the rest of the articles dealing with biosurveillance.

1 Evaluation

There is a growing body of literature on evaluation of "syndromic surveillance" that ranges broadly. Topics include

1. Advice on what to consider as a framework [CDC04]
2. Assessment of specific data source validity [FSS04]
3. Algorithm performance [MRC04, SD03, BBC05]
4. General policy discourse [Rei03, SSM04]
5. Activity overview, etc. [BBC05]

The breadth of this subtopic attests to the interest of evaluation for a relatively immature area in public health surveillance system development. It does not seem reasonable to expect meaningful evaluative conclusions about surveillance systems (e.g., cost/benefit utility) without a means to rigorously evaluate system *components* individually.

Consider that for modern biosurveillance systems, there are

1. Information technology process segments for recording electronic transactions and moving data.
2. Data preprocessing functions that include structuring an accessible analytic database architecture and ensuring data quality.
3. Data analysis components to apply methods for inference as well as deduction.
4. Support tools that operate in a decision theoretic framework for combining evidence, other information, and communication to facilitate action in near realtime.

To acquire useful evaluation measures for a surveillance system, subcategories are required so that specific enough objectives could be established. By this approach, evaluation for the provincial notion of "whether or not" to do surveillance gets replaced with the more practical notion of "how to do it better." Also, system complexity is reduced by decomposition. A risk here is to oversegregate interdependent activities and create operational stovepipes among professional skill sets. Good management and leadership must be alert and proactive to prevent maladaptive marginalization of system development, data management, statistical subject matter or end-user professionals in evaluation research and subsequent development activities for surveillance. This is essential in order to "conquer" after dividing; or more specifically in this context, to make sense out of algorithm performance characteristics after considering them separately from other operational surveillance components.

Since much of the data used for public health surveillance are not collected specifically for that purpose and/or are spontaneously generated, (1) they are referred to as "secondary" or "opportunistic," (2) the data require substantial preprocessing for analytic use, (3) a sample-to-population mapping is not probabilistically defined, and therefore, (4) the analytic signal detection

methodologies are empirical in nature and do not lend themselves to conclusions bearing well-defined inferential quantities such as confidence intervals or p-values. There is generally no sampling design to define the probabilistic relationship between the data and a specified population of interest, and a design-based guide for an analytic strategy in a traditional sense is absent. Therefore effective use of these systems is primarily empirical.

Detection algorithm performance evaluation in an empirical setting is problematic when events for detection are rare. In the absence of recorded events of importance to train upon, thoughtful and informed simulation is much needed to accelerate learning. Ideally, a realistically described scenario can be translated into representation in data as a response to people's behavior. Characteristics of the scenario that would affect representation in the data could be modified with a consequential data representation. Monte Carlo iterations of the simulated signal structured over real data absent of events of interest could then be cycled with detection activities recorded. Thus, the usual means to evaluate a statistical detection approach for its operating characteristics under varying conditions could be established. What frequently takes place is that the people or groups who develop and promote a detection approach are the same ones who establish the simulation and the evaluation criteria and interpret the outcome. This is certainly a reasonable first step but this process leaves too much opportunity for scientific confounding — designing the evaluation criteria to fit the object of evaluation. A more objective approach would serve to advance the field more effectively.

In addition to (1) well-defined signals of importance, (2) the use of simulation, and (3) increased objectivity, the results of evaluation studies for surveillance system performance are of much greater practical value if they consider the realistic operational conditions under which data analysts must make decisions. Three considerably influential factors are data "lag time," "time alignment," and the "unlinked multiple data source" problem. Two ways that data lag time can be considered are (1) the average time between a population event (e.g., patient encounter or some other health-seeking behavioral event) and the event's data representation in an analytic system interface or (2) the proportion of data available at the time a decision is needed (versus at some later time). "Time alignment" refers to the differential health-seeking behavior times relevant for various data sources that may be available in one analytic system. For example, if one were able to view time series signals in response to a population exposure that caused illness, it may appear earlier for sales data than for emergency department (ED) data. The reasoning is that people may generally purchase products for self-treatment before their symptoms would be severe enough to warrant a trip to the ED. The "unlinked data source" problem is an issue for the secondary use of data sources when record linkage is either not possible or avoided for other reasons. Given that much of the data used in automated surveillance is gathered for some other purpose (treating patients, billing, market analysis, inventory, etc.) and that protecting individual confidentiality is a motive, broad linkage of records is

not generally feasible. Therefore, the extent of information overlap is unknown across data streams. For example, if a system uses over-the-counter sales, ED, and laboratory test order data, it is not known to what extent the same people and their reactions to illness are manifest in the different sources. Without consideration of these operational realities, simulations for determining operating characteristics of new surveillance paradigms are incomplete at best and of marginal practical value.

2 Coordination for Information Exchange among Jurisdictions: BioSense

This is an aspect of analyzing and using information that easily goes unnoticed or is not well understood by the technical data analysis professionals who develop the analytic methodologies of surveillance systems. In public health as well as many public service industries, local jurisdictions are the primary users of information systems relating to situational awareness and their potential need to respond in their communities. When situations cross jurisdictional borders, coordinating response becomes a shared challenge. When public health threats cross state borders, the federal government becomes responsible for coordinating information. The time and efficiency of meeting this challenge are facilitated greatly through the use of technology standards [Bra05]. Conversely, multiple and diverse system outputs are difficult to exchange and consequently interpret. Thus, considering the potential public health threat that bioterrorism poses, there is a critical need for standards in data coding and preprocessing, data management procedures, analytic algorithms, data monitor operating procedures, and documentation of anomaly investigations. Further, since it could be any part of the nation that is at risk, these standards need to be national in scope. The Centers for Disease Control and Prevention has launched an initiative called BioSense to serve as a platform for standards development as part of the Public Health Information Network [Loo04]. BioSense is intended to provide a national safety net ensuring that early detection is enabled in all major metropolitan areas and works to support and integrate with existing regional surveillance systems. Requirements, data characteristics, threshold tolerances for response potential, etc. will likely continue to be different among local areas but it is certainly in our interests to enable rapid exchange of analytic results across jurisdictions. The goal is to have standard statistical and other data analytic conceptual approaches that can be tailored to local needs using various user-defined settings, results from which can be described coherently in a way that provides interoperable information for national situation awareness.

3 An Open Issue: Null Hypothesis Dilemma

An open question that is worth consideration in both advancing probabilistic methods for surveillance data analysis relates to the type I and type II error concepts. If we consider the null condition to be the assumption of "no event of importance in progress" and the alternative to be supported when there is sufficient data to conclude that a countermeasure response is needed, then the type I error is defined to be falsely concluding that a response is needed when in fact it is not necessary. This seems like the less important "mistake" in that if something were occurring that warranted a reaction and we did not respond, lives would be lost and precious time would have passed in stopping an event of importance. Thus, our general approach to controlling the type I error using "alpha" for threshold setting is questionable in this setting. On the other hand, being overly conservative at the expense of allowing too many false alerts may fatigue readiness resulting in an inability to respond when truly needed. The goal is to strike an informed balance between sensitivity maintenance and false alert toleration. Currently implemented surveillance systems in public health are based on inferential concepts that use p-values for thresholds under the null assumption that the situation is expected with relation to the temporal and/or geographical context. Given the situational consequences of failing to alert to true events and too frequently alerting to unimportant events, more refined bases for conclusions must be established as standard operating procedures using decision theoretic approaches and specifying risk and utility functions.

4 Summary and Directions

What has been commonly referred to as "syndromic surveillance" is not well-defined and is quickly growing out of its previous characterization. The implementation of new operational models for early event detection and subsequent situational awareness is creating opportunities for statistical and other data analytic applications in public health. Challenges include the following:

- There is little collective working experience with secondary data use among analysts.
- Data systems are new relative to the statistical methodologies employed.
- Data management tasks are large and the human resource skill sets for accomplishing those tasks are rare and underrated.
- Successful information system operations require close communications among staff of several interdependent disciplines.
- Analysis of these data requires inductive and deductive reasoning in combination (results may be difficult to communicate concisely).
- Multiple data streams:
  1. How can we best approach analysis: multiunivariate or multivariate?

2. There is a knowledge gap for population behavioral response patterns (the time alignment question).

The practice of binning population events into categories of likely association with syndromes relating to known serious biological agents, counting, comparing, and looking for patterns is currently the basis for most of the work in this area. This seems a logical first iteration of maturity for a surveillance system to enable earlier detection than would be possible otherwise. There is a need to apply decision science concepts to support end-user's threshold determination. The use of prior knowledge in a Bayesian framework and more refined pattern recognition seems like a promising direction for detection refinement, especially as more detailed data can be consolidated and means to process it are built. As more diverse data sources are integrated (human health, animal health, plant health, water quality, Internet traffic, utilities, intelligence, etc.), analytic approaches and applied methodologies for combining evidence from multiple and often conflicting sources will become even more important [SF02]. In the meantime, simulation appears to be the most promising method for accelerating available working knowledge of empirical surveillance.

In the chapters of Part III that follow, Shmueli and Fienberg provide an informed listing and brief conceptual characterization for a spectrum of detection approaches that either have already been implemented or hold promise for utility in surveillance. Their attention is primarily on the statistical methodologies and use of data from multiple sources, a logical focus given the current state of systems in application. Stoto et al. continue in this topic by creatively comparing the empirical detection performance of algorithms using simulated changes in patterns embedded in real health care data from Washington, DC. Finally, Forsberg et al. develop in an elegant historical context, the elucidation of how to take advantage of the space and time dimensions simultaneously in identifying clusters of events.

References

[Bra05] Bradwell, W. R. 2005. Enterprise architecture for health departments. Public Health Informatics Institute, http://www.phii.org/pages/Bradwell.html.

[BBC05] Buckeridge, D. L., H. Burkom, M. Campbell, W. R. Hogan, and A. W. Moore. 2005. "Algorithms for rapid outbreak detection: A research synthesis." *Journal of Biomedical Informatics* 38:99–113.

[CDC04] Centers for Disease Control and Prevention. 2004. "Framework for evaluating public health surveillance systems for early detection of outbreaks; recommendations from the CDC Working Group." *Morbidity and Mortality Weekly Report* 53 (RR-5): 1–13.

[FSS04] Fleischauer, A. T., B. J. Silk, M. Schumacher, K. Komatsu, S. Santana, V. Vaz, M. Wolfe, L. Hutwagner, J. Cono, R. Berkelman, and T. Treadwell. 2004. "The validity of chief complaint and discharge diagnosis in

emergency department-based syndromic surveillance." *Academic Emergency Medicine Journal* 11 (12): 1262–1267 (December).

[Hen04] Henning, K. J. 2004. "What is syndromic surveillance?" *Morbidity and Mortality Weekly Report* 53 (Supplement): 7–11. Syndromic Surveillance: Reports from a National Conference, 2003.

[Loo04] Loonsk, J. W. 2004. "BioSense — A national initiative for early detection and quantification of public health emergencies." *Morbidity and Mortality Weekly Report* 53 (Supplement): 53–55. Syndromic Surveillance: Reports from a National Conference, 2003.

[MRC04] Mandl, K. D., B. Reis, and C. Cassa. 2004. "Measuring outbreak-detection performance by using controlled feature set simulations." *Morbidity and Mortality Weekly Report* 53 (Supplement): 130–136. Syndromic Surveillance: Reports from a National Conference, 2003.

[MH03] Mostashari, F., and J. Hartman. 2003. "Syndromic surveillance: A local perspective." *Journal of Urban Health* 80 (Suppl. 1): i1–i7.

[Rei03] Reingold, A. 2003. "If syndromic surveillance is the answer, what is the question?" *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 1 (2): 1–5.

[SF02] Sentz, K., and S. Ferson. 2002. "Combination of evidence in Dempster-Shafer theory." Sandia report SAND2002-0835, Sandia National Laboratories, Albuquerque, NM.

[SD03] Sosin, D. M., and J. DeThomasis. 2004. "Evaluation challenges for syndromic surveillance — making incremental progress." *Morbidity and Mortality Weekly Report* 53 (Supplement): 125–129. Syndromic Surveillance: Reports from a National Conference, 2003.

[SSM04] Stoto, M. A., M. Schonlau, and L. T. Mariano. 2004. "Syndromic surveillance: Is it worth the effort?" *Chance* 17:19–24.

Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Biosurveillance

Galit Shmueli[1] and Stephen E. Fienberg[2]

[1] Decision and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, gshmueli@rhsmith.umd.edu
[2] Department of Statistics, The Center for Automated Learning and Discovery, and Cylab, Carnegie Mellon University, fienberg@stat.cmu.edu

1 Introduction

A recent review of the literature on surveillance systems revealed an enormous number of research-related articles, a host of websites, and a relatively small (but rapidly increasing) number of actual surveillance systems, especially for the early detection of a bioterrorist attack [BMS04]. Modern bioterrorism surveillance systems such as those deployed in New York City, western Pennsylvania, Atlanta, and Washington, DC, routinely collect data from multiple sources, both traditional and nontraditional, with the dual goal of the rapid detection of localized bioterrorist attacks and related infectious diseases. There is an intuitive notion underlying such detection systems, namely, that detecting an outbreak early enough would enable public health and medical systems to react in a timely fashion and thus save many lives. Demonstrating the real efficacy of such systems, however, remains a challenge that has yet to be met, and several authors and analysts have questioned their value (e.g., see Reingold [Rei03] and Stoto et al. (2004) [SSM04, SFJ06]). This article explores the potential and initial evidence adduced in support of such systems and describes some of what seems to be emerging as relevant statistical methodology to be employed in them.

Public health and medical data sources include mortality rates, lab results, emergency room (ER) visits, school absences, veterinary reports, and 911 calls. Such data are directly related to the treatment and diagnosis that would follow a bioterrorist attack. They might not, however, detect the outbreak sufficiently fast. Several recent national efforts have been focused on monitoring "earlier" data sources for the detection of bioterrorist attacks or other outbreaks, such as over-the-counter (OTC) medication sales, nurse hotlines, or even searches on medical websites (e.g., WebMD). This assumes that people who are not aware of the outbreak and are feeling sick, would gen-

erally seek self-treatment before approaching the medical system and that an outbreak signature will manifest itself earlier in such data. According to Wagner et al. [WRT03], preliminary studies suggest that sales of OTC health care products can be used for the early detection of outbreaks, but research progress has been slow due to the difficulty that investigators have in acquiring suitable data to test this hypothesis for sizable outbreaks. Some data of this sort are already being collected (e.g., pharmacy and grocery sales). Other potential nontraditional data sources that are currently not collected (e.g., browsing in medical websites, automatic body sensor devices) could contain even earlier signatures of an outbreak.[3]

To achieve rapid detection there are several requirements that a surveillance system must satisfy: frequent data collection, fast data transfer (electronic reporting), real-time analysis of incoming data, and immediate reporting. Since the goal is to detect a large, localized bioterrorist attack, the collected information must be local, but sufficiently large to contain a detectable signal. Of course, the different sources must carry an early signal of the attack. There are, however, trade-offs between these features; although we require frequent data for rapid detection, too frequent data might be too noisy to the degree that the signal is too weak for detection. A typical solution for too frequent data is temporal aggregation. Two examples where aggregation is used for biosurveillance are aggregating OTC medication sales from hourly to daily counts [GSC02] and aggregating daily hospital visits into multiday counts [RPM03]. A similar trade-off occurs between the level of localization of the data and their amount. If the data are too localized, there might be insufficient data for detection, whereas spatial aggregation might dampen the signal.

Another important set of considerations that limit the frequency and locality of collected data relate to confidentiality and data disclosure issues (concerns over ownership, agreements with retailers, personal and organizational privacy, etc.). Finding a level of aggregation that contains a strong enough signal, that is readily available for collection without confronting legal obstacles, and yet is sufficiently rapid and localized for rapid detection, is clearly a challenge. We describe some of the confidentiality and privacy issues briefly here.

There are many additional challenges associated with the phases of data collection, storage, and transfer. These include standardization, quality control, confidentiality, etc. [FS05]. In this paper we focus on the statistical challenges associated with the data monitoring phase, and in particular, data in the form of multiple time series. We start by describing data sources that are

[3] While our focus in this article is on passive data collection systems for syndromic surveillance, there are other active approaches that have been suggested (e.g., screening of blood donors [KPF03]), as well as more technological fixes, such as biosensors [Sul03] and "Zebra" chips for clinical medical diagnostic recording, data analysis, and transmission [Cas04].

collected by some major surveillance systems and their characteristics. We then examine various traditional monitoring tools and approaches that have been in use in statistics in general, and in biosurveillance in particular. We discuss their assumptions and evaluate their strengths and weaknesses in the context of biosurveillance. The evaluation criteria are based on the requirements of an automated, nearly real-time surveillance system that performs online (or prospective) monitoring of incoming data. These are clearly different than for retrospective analysis [SB03] and include computational complexity, ease of interpretation, roll-forward features, and flexibility for different types of data and outbreaks.

Currently, the most advanced surveillance systems routinely collect data from multiple sources on multiple data streams. Most of the actual statistical monitoring, however, is typically done at the univariate time series level, using a wide array of statistical prediction methodologies. Ideally, multivariate methods should be used so that the data can be treated in an integrated way, accounting for the relationships between the data sources. We describe the traditional statistical methods for multivariate monitoring and their shortcomings in the context of biosurveillance. Finally, we describe monitoring methods, in both the univariate and multivariate sections, that have evolved in other fields and appear potentially useful for biosurveillance of traditional and nontraditional temporal data. We describe the methods and describe their strengths and weaknesses for modern biosurveillance.

2 Types of Data Collected in Surveillance Systems

Several surveillance systems aimed at rapid detection of disease outbreaks and bioterror attacks have been deployed across the United States in the last few years, including the *Realtime Outbreak and Disease Surveillance* system (RODS) and *National Retail Data Monitor* (NRDM) in western Pennsylvania, the *Early Notification of Community-Based Epidemics system* (ESSENCE) in the Washington, DC, area (which also monitors many Army, Navy, Air Force, and Coast Guard data worldwide), the *New York City Department of Health and Mental Hygiene* (NYC-DOHMH) system, and recently the *BioSense* system by the Centers for Disease Control and Prevention. Each system collects information on multiple data sources with the intent of increasing the certainty of a true alarm by verifying anomalies found in various data sources [PMK03]. All of these systems collect data from medical facilities, usually at a daily frequency. These include emergency rooms admissions (RODS, NYC-DOHMH), visits to military treatment facilities (ESSENCE), and 911 calls (NYC-DOHMH). Nontraditional data include OTC medication and healthcare product sales at grocery stores and pharmacies (NRDM, NYC-DOHMH, ESSENCE), prescription medication sales (ESSENCE), HMO billing data (ESSENCE), and school/work absenteeism records (ESSENCE). We can think of the data in a hierarchical structure; the first level consists of the data source

(e.g., ER or pharmacy), and then within each data source there might be multiple time series, as illustrated in Fig. 1.

This structure suggests that series that come from the same source should be more similar to each other than to series from different sources. This can influence the type of monitoring methods used within a source as opposed to methods for monitoring the entire system. For instance, within-source series will tend to share variation sources such as holidays, closing dates, and seasonal effects. Pharmacy holiday closing hours will influence all medication categories equally but not school absences. From a modeling point of view this structure raises the question whether a hierarchical model is needed or else all series can be monitored using a flat multivariate model. In practice, most traditional multivariate monitoring schemes and a wide range of applications consider similar data streams. Very flexible methods are needed to integrate all the data within a system that is automatic, computationally efficient, timely, and with low false alarms. In the following sections we describe univariate and multivariate methods that are currently used or can potentially be used for monitoring the various multiple data streams. We organize and group the different methods by their original or main field of application and discuss their assumptions, strengths, and limitations in the context of biosurveillance data.

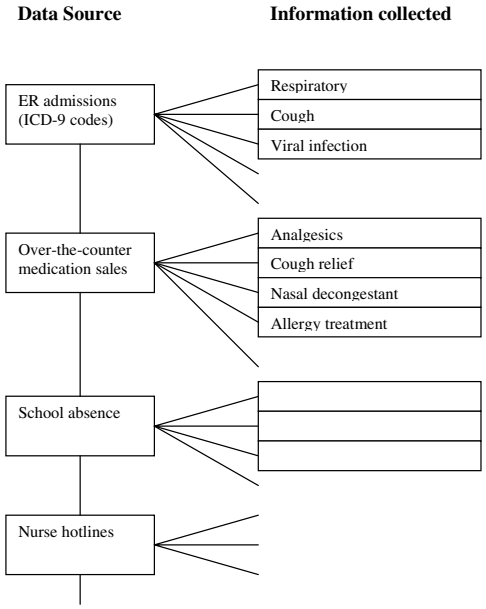

Fig. 1. Sketch of data hierarchy; each data source can contain multiple time series.

3 Monitoring Univariate Data Streams

The methods used in biosurveillance borrow from several fields, with statistical process control being the most influential. Methods from other fields have also been used occasionally, with most relying on traditional statistical methods such as regression and time series models. Although different methods might be more suitable for different data streams or sources, there are advantages to using a small set of methods for all data streams within a single surveillance system. This simplifies automation, interpretability, and coherence, and the ability to integrate results from multiple univariate outputs. The principle of parsimony, which balances performance and simplicity, should be the guideline.

We start by evaluating some of the commonly used monitoring methods and then describe other methods that have evolved or have been applied in other fields, which are potentially useful for biosurveillance.

3.1 Current Common Approaches

Statistical Process Control

Monitoring is central to the field of statistical process control. Deming, Shewhart, and others revolutionized the field by introducing the need and tools for monitoring a process to detect abnormalities at the early stages of production. Since the 1920s the use of control charts has permeated into many other fields including the service industry. One of the central tools for process control is the control chart, which is used for monitoring a parameter of a distribution. In its simplest form the chart consists of a centerline, which reflects the target of the monitored parameter and control limits. A statistic is computed from an *iid* sample every time point, and its value is plotted on the chart. If it exceeds the control limits, the chart flags an alarm, indicating a change in the monitored parameter. Statistical methods for monitoring univariate and multivariate time series tend to be model-based. The most widely used control charts are Shewhart charts, moving average (MA) charts, and cumulative sum (CuSum) charts. Each of these methods specializes in detecting a particular type of change in the monitored parameter [BL97].

We now briefly describe the different charts. Let $\mathbf{y}_t$ be a random sample of measurements taken at time t ($t = 1, 2, 3, \ldots$). In a Shewhart chart the monitoring statistic at time t, denoted by S_t, is a function of $\mathbf{y}_t$:

$$S_t = f(\mathbf{y}_t). \tag{1}$$

The statistic of choice depends on the parameter that is monitored. For instance, if the process mean is monitored, then the sample mean ($f(\mathbf{y}_t) = \overline{\mathbf{y}}_t$) is used. If the process variation is monitored, a popular choice is the sample standard deviation ($f(\mathbf{y}_t) = s_t$). The monitoring statistic is drawn on a time

plot, with lower and upper control limits. When a new sample is taken, the point is plotted on the chart, and if it exceeds the control limits, it raises an alarm. The assumption behind the classic Shewhart chart is that the monitoring statistic follows a normal distribution. This is reasonable when the sample size is large enough relative to the skewness of the distribution of $\mathbf{y}_t$. Based on this assumption, the control limits are commonly selected as ± 3 standard deviations of the monitoring statistic (e.g., if the sample mean is the monitoring statistic, then the control limits are $\pm 3\sigma/\sqrt{n}$) to achieve a low, false-alarm rate of $2\phi(-3) = 0.0027$. Of course, the control limits can be chosen differently to achieve a different false-alarm rate. If the sample size at each time point is $n = 1$, then we must assume that $\mathbf{y}_t$ are normally distributed for the chart to yield valid results. Alternatively, if the distribution of $f(\mathbf{y}_t)$ (or $\mathbf{y}_t$) is known, then a valid Shewhart chart can be constructed by choosing the appropriate percentiles of that distribution for the control limits as discussed in [SFJ06].

Shewhart charts are very popular because of their simplicity. They are very efficient at detecting moderate-to-large, spike-type changes in the monitored parameter. Since they do not have a "memory," a large spike is immediately detected by exceeding the control limits. However, Shewhart charts are not useful for detecting small spikes or longer-term changes. In those instances we need to retain a longer "memory." One solution is to use the "Western Electric" rules. These rules raise an alarm when a few points in a row are too close to a control limit, even if they do not exceed it. Although such rules are popular and are imbedded in many software programs, their addition improves detection of real aberrations at the cost of increased false alarms. The trade-off turns out to be between the expected time-to-signal and its variability [SC03].

An alternative is to use statistics that have longer memories. Three such statistics are the MA, the exponentially weighted moving average (EWMA), and the CuSum. MA charts use a statistic that relies on a constant-size window of the k last observations:

$$\mathrm{MA}_t = \sum_{j=1}^{k} f(\mathbf{y}_{t-j+1})/k. \tag{2}$$

The most popular statistic is a grand mean ($\sum_{j=1}^{k} \overline{y}_{t-j+1}/k$). These charts are most efficient for detecting a step increase/decrease that lasts k time points.

The original CuSum statistic defined by $\frac{1}{\sigma}\sum_{i=1}^{t}(y_t - \mu)$ keeps track of all the data until time t [HO98]. However, charts based on this statistic are awkward graphically. A widely used adaptation is the tabular CuSum, which restarts the statistic whenever it exceeds zero. The one-sided tabular CuSum for detecting an increase is defined as

$$\mathrm{CuSum}_t = \max\{0, (y_t^* - k) + \mathrm{CuSum}_{t-1}\}, \tag{3}$$

where $y_t^* = (y_t - \mu)/\sigma$ are the standardized observations, and k is proportional to the size of the abnormality that we want to detect. This is the most efficient

statistic for detecting a step change in the monitored parameter. However, it is less useful for detecting a spike since it would be masked by the long memory. In general, time series methods that place heavier weight on recent values are more suitable for short-term forecasts [Arm01].

The EWMA statistic is similar to the CuSum, except that it weights the observations as a function of their recency, with recent observations taking the highest weight:

$$\text{EWMA}_t = \alpha y_t + (1-\alpha)\text{EWMA}_{t-1} = \alpha \sum_{j=0}^{t-1} (1-\alpha)^j f(y_{t-j}) + (1-\alpha)\text{EWMA}_0, \tag{4}$$

where $0 < \alpha \leq 1$ is the smoothing constant [Mon01]. This statistic is best at detecting an exponential increase in the monitored parameter. It is also directly related to exponential smoothing methods (see below). For further details on these methods, see Montgomery [Mon01]. In biosurveillance, the EWMA chart was used for monitoring weekly sales of OTC electrolytes to detect pediatric respiratory and diarrheal outbreaks [HTI03] and is used in ESSENCE II to monitor ER admissions in small geographic regions [Bur03a].

Since the statistic in these last three cases is a weighted average/sum over time, the normality assumption of y_t is less crucial for adequate performance due to the central limit theorem, especially in the case of the EWMA [RS04b, ACV04]. The main disadvantage of all these monitoring tools is that they assume statistical independence of the observations. Their original and most popular use is in industrial process control where samples are taken from the production line at regular intervals, and a statistic based on these assumably independent samples is computed and drawn on the control chart. The *iid* assumption is made in most industrial applications, whether correct or not. Sometimes the time between samples is increased to minimize correlation between close samples. In comparison, the types of data collected for biosurveillance are usually time series that are collected on a frequent basis to achieve timeliness of detection, and therefore autocorrelation is inherent. For such dependent data streams the use of distribution-based or distribution-free control charts can be misleading in the direction of increased false-alarm rates [Mon01, p. 375].

A common approach to dealing with autocorrelated measurements is to approximate them using a time series model and monitor the residual error using a control chart [GDV04]. The assumption is that the model accounts for the dependence and therefore the residuals should be nearly independent. Such residuals will almost never be completely independent, however, and the use of control charts to monitor them should be done cautiously. This is where time series analysis emerges in anomaly detection applications in general, and in biosurveillance in particular. Moreover, because the forecast at every time point is used to "test" for anomalies, we need to deal with the multiple testing problem for dependent tests and possibly use variations on

the new literature on false discovery rates (FDR) to control familywise type I errors [BH95, EST01].

Time Series Methods

The most well-known class of time series models used by statisticians is autoregressive moving average (ARMA) models. Conceptually they are similar to regressing the current observations on a window of previous observations while assuming a particular autocovariance structure. An ARMA(p,q) model is defined as

$$y_t = \mu + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \tag{5}$$

where α_i and θ_j are parameters and $\varepsilon_{t-q} \ldots \varepsilon_t$ are white noise (having mean 0 and standard deviation σ_ε). To fit an ARMA model, the modeler must determine the order of the autoregression p and the order of the MA component, q. This task is not straightforward and requires experience and expertise (for example, not every selection of p and q yields a causal model). After p and q are determined, there are $p + q + 1$ parameters to estimate, usually through nonlinear least squares (LS) and conditional maximum likelihood. The process of selecting p and q and estimating the parameters is cyclical [BJR94] and typically takes several cycles until a satisfactory model is found. Some software packages do exist that have automated procedures for determining p and q and estimating those parameters.

ARMA models can combine external information by adding predictors in the model. This allows to control for particular time points that are known to have a different mean by adding indicators with those time points. Such modifications are especially useful in the biosurveillance context, since effects such as weekend/weekday and holidays are normally present in medical and nontraditional data. ESSENCE II, for instance, uses an autoregressive model that controls for weekends, holidays, and postholidays through predictors [Bur03a].

ARMA models assume that the series is stationary over time (i.e., the mean, variance, and autocovariance of the series remain constant throughout the period of interest). In practice, fitting of an ARMA model to data usually requires an initial preprocessing step where the data are transformed in one or more ways until a stationary or approximately stationary series emerges. The most popular generalization of ARMA models for handling seasonality and trends is to add a differencing transformation, thereby yielding an autoregressive integrated moving average (ARIMA) model of the form

$$(1-\alpha_1 B-\alpha_2 B^2-\ldots-\alpha_p B^p)[(1-B)^d(1-Bs)^D y_t-\mu] = (1-\theta_1 B-\ldots-\theta_q B^q)\varepsilon_t, \tag{6}$$

where B is the back-shift operator ($By_t = y_{t-1}$), $d > 0$ is the degree of nonseasonal differencing, $D > 0$ is the degree of seasonal differencing, and s is the length of a seasonal cycle. Determining the level of differencing is

not trivial, and over- and underdifferencing can lead to problems in modeling and forecasting [CR96]. Although this model allows flexibility, in practice the model identification step is complicated and highly data specific, and requires expertise of the modeler. Another disadvantage of ARIMA models is their computational complexity. With thousands of observations, the method requires considerable computer time and memory [SAS04b].

To summarize, the common statistical approach towards monitoring has been mostly distribution based. Recent advances in data availability and collection in the process industry have led authors such as Willemain and Runger [WR96] to emphasize the importance of model-free methods. It appears, though, that such methods have already evolved and have been used in other fields! Next, we describe a few such methods that are distribution-free.

3.2 Monitoring Approaches in Other Fields

Monitoring methods have been developed and used in other fields such as machine learning, computer science, geophysics, and chemical engineering. Also, forecasting, which is related to monitoring, has had advances in fields such as finance and economics. In these fields there exist a wealth of very frequent autocorrelated data; the goal is the rapid detection of abnormalities ("anomaly detection") or forecasting, and the developed algorithms are flexible and computationally efficient. We describe a few of the methods used in these fields and evaluate their usefulness for biosurveillance.

Anomaly detection in machine learning emphasizes automated and usually model-free algorithms that are designed to detect local abnormalities. Even within the class of model-free algorithms, there is a continuum between those that are intended to be completely "user-independent" and those that require expert knowledge integration by the user. An example for the former is the symbolic aggregate approximation (SAX), which is a symbolic representation for time series that allows for dimensionality reduction and indexing [LKL03]. According to its creators, "anomaly detection algorithms should have as few parameters as possible, ideally none. A parameter free algorithm would prevent us from imposing our prejudices, expectations, and presumptions on the problem at hand, and would let the data itself speak to us" [KLR04]. In biosurveillance there exists expert knowledge about the progress of a disease, its manifestation in medical and public health data, etc. An optimal method would then be distribution-free and parsimonious, but would allow the integration of expert knowledge in a simple way.

Exponential Smoothing

Exponential smoothing (ES) is a class of methods that is very widely used in practice (e.g., for production planning, inventory control, and marketing [PA89]) but not so in the biosurveillance field. ES has gained popularity mostly because of its usefulness as a short-term forecasting tool. Empirical research by

Makridakis et al. [MAC82] has shown simple exponential smoothing (SES) to be the best choice for one-step-ahead forecasting, from among 24 other time series methods and using a variety of accuracy measures. Although the goal in biosurveillance is not forecasting, ES methods are relevant because they can be formulated as models [CKO01]. Nontraditional biosurveillance data include economic series such as sales of medications, health-care products, and grocery items. Since trends, cycles, and seasonality are normally present in sales data, more advanced ES models have been developed to accommodate nonstationary time series with additive multiplicative seasonality/linear/exponential/dampened trend components. A general formulation of an ES model assumes that the series is comprised of a level, trend (the change in level from last period), seasonality (with M seasons), and error. To illustrate the model formulation, estimation, and forecasting processes, consider an additive model of the form

$$y_t = \text{local mean} + \text{seasonal factor} + \text{error}, \tag{7}$$

where the local mean is assumed to have an additive trend term and the error is assumed to have zero mean and constant variance. At each time t, the smoothing model estimates these time-varying components with level, trend, and seasonal smoothing states denoted by L_t, T_t, and $S_{t-i}(i = 0, \ldots, M-1)$, respectively.[4] The smoothing process starts with an initial estimate of the smoothing state, which is subsequently updated for each observation using the *updating equations*:

$$\begin{aligned} L_{t+1} &= \alpha(y_{t+1} - S_{t+1-M}) + (1-\alpha)(L_t + T_t), \\ T_{t+1} &= \beta(L_{t+1} - L_t) + (1-\beta)T_t, \\ S_{t+1} &= \gamma(y_{t+1} - L_{t+1}) + (1-\gamma)S_{t+1-M}, \end{aligned} \tag{8}$$

where α, β, and γ are the smoothing constants. The m-step-ahead forecast at time t is

$$\hat{y}_{t+m} = L_t + mT_t + S_{t+m-M}. \tag{9}$$

A multiplicative model of the form $Y_t = (L_{t-1} + tT_{t-1})S_{t-i}\,\varepsilon_t$ can be obtained by applying the updating equations in (8) to $\log(y_t)$. The initial values L_0, T_0, and the M seasonal components at time 0 can be estimated from the data using a centered MA (see Pfeffermann and Allon [PA89] and the NIST Handbook [NIS04] for details). The three smoothing constants are either determined by expert knowledge, or estimated from the data by maximizing a well-defined loss function (e.g., mean of squared one-step-ahead forecast errors).

From a modeling point of view, many ES methods have ARIMA, seasonal ARIMA (SARIMA), and structural models equivalents, and they even include a class of dynamic nonlinear state space models that allow for changing

[4] The smoothing state is normalized so that the seasonal factors S_{t-i} for $i = 0, 1, \ldots, M$ sum to zero for models that assume additive seasonality, and average to one for models that assume multiplicative seasonality [CY88].

variance [CKO01]. Table 1 summarizes some of these equivalences. It is noteworthy that some of the SARIMA equivalents are so complicated that they are most unlikely to be identified in practice [CKO01]. Furthermore, Chatfield et al. [CKO01] show that there are multiple generating processes for which a particular ES method is optimal in the sense of forecast accuracy, which explains their robust nature. The advantage of these models is their simplicity of implementation and interpretation, their flexibility for handling many types of series, and their suitability for automation [CY88] because of the small number of parameters involved and the low computational complexity. They are widely used and have proved empirically useful, and automated versions of them are available in major software packages such as the high-performance forecasting module by SAS®[SAS04a].

Table 1. The equivalence between some exponential smoothing and (seasonal)-ARIMA models. The notation $ARIMA(p,d,q)(P,D,Q)_s$ corresponds to an ARIMA(p,d,q) with seasonal cycle of length s, P-order autoregressive seasonality, seasonal differencing of order D, and seasonal MA of order Q

Exponential Smoothing Method	ARIMA/SARIMA Equivalent
Simple exponential smoothing	ARIMA(0,1,1)
Holt's (double) linear trend method	ARIMA(0,2,2)
Damped-trend linear method	ARIMA(1,1,2)
Additive Holt-Winters (triple) method	SARIMA(0,1,p+1)$(0,1,0)_p$
Multiplicative Holt-Winters (triple) method	[KSO01]'s dynamic nonlinear state-space models

Singular Spectral Analysis

The methods of singular spectral analysis (SSA) were developed in the geosciences as an alternative for Fourier/spectral analysis and have been used mostly for modeling climatic time series such as global surface temperature records [GV91], and the Southern Oscillation Index that is related to the recurring El Niño/Southern Oscillations conditions in the Tropical Pacific [PGV95, YSG00] . It is suitable for decomposing a short, noisy time series into a (variable) trend, periodic oscillations, other aperiodic components, and noise [PGV95].

SSA is based on an eigenvalue-eigenvector decomposition of the estimated M-lag autocorrelation matrix of a time series, using a Karhunen-Loeve decomposition. The eigenvectors, denoted by $\varrho_1, \ldots, \varrho_M$, are called empirical orthogonal functions (EOFs) and form an optimal basis that is orthonormal at lag zero. Usually a single EOF is sufficient to capture a nonlinear oscillation. Using statistical terminology, principal components analysis (PCA) is applied to the estimated autocorrelation matrix, so that projecting the EOFs on the time series gives the principal components $(\Lambda_1, \ldots, \Lambda_M)$:

$$\Lambda_k(t) = \sum_{i=1}^{M} y(t+i)\varrho_k(i), \tag{10}$$

and the eigenvalues reflect the variability associated with the principal components [GY96]. The next step in SSA is to reconstruct the time series using only a subset $\mathcal{K}$ of the EOFs:

$$y_{\mathcal{K}}(t) = \frac{1}{M_t} \sum_{k \in \mathcal{K}} \sum_{i=1}^{M} \Lambda_k(t-i)\varrho_k(i), \tag{11}$$

where M_t is a normalizing constant (for details, see [GV91]). Choosing $\mathcal{K}$ is done heuristically or by Monte Carlo simulations.

SSA is used mostly for revealing the underlying components of a time series and separating signal from noise. However, it can be used for forecasting by using low-order autoregressive models for the separate reconstructed series [PGV95]. This means that SSA can be used for biosurveillance and monitoring in general by computing one-step-ahead forecasts and comparing them to the actual data. If the distance between a forecast and an actual observation is too large, a signal is triggered.

Although SSA assumes stationarity (by decomposing the estimated autocorrelation matrix), according to Yiou et al. [YSG00], it appears less sensitive to nonstationarity than spectral analysis. However, Yiou et al. [YSG00] suggested a combination of SSA with wavelets to form multiscale SSA (MS-SSA). The idea is to use the EOFs in a data-adaptive fashion with a varying window width, which is set as a multiple of the order M of the autocorrelation matrix. After applying the method to synthetic and real data, they conclude that MS-SSA behaves similarly to wavelet analysis, but in some cases it provides clearer insights into the data structure. From a computational point of view, MS-SSA is very computationally intensive, and in practice a subset of window widths is selected rather than exhaustively computing over all window widths [YSG00].

Wavelet-Based Methods

An alternative to ARIMA models that has gained momentum in the last several years is wavelet decomposition. The idea is to decompose the time series $y(t)$ using wavelet functions:

$$y(t) = \sum_{k=1}^{N} a_k \phi(t-k) + \sum_{k=1}^{N} \sum_{j=1}^{m} d_{j,k} \psi(2^j t - k), \tag{12}$$

where a_k is the scaled signal at time k at the coarsest scale m, $d_{j,k}$ is the detail coefficient at time k at scale j, ψ is a scaling function (known as the "father wavelet"), and ϕ is the mother wavelet function.

This method is very useful in practice, since data from most processes are multiscale in nature due to "events occurring at different locations and with different localization in time and frequency, stochastic processes whose energy or power spectrum changes with time and/or frequency, and variables measured at different sampling rates" [Bak98]. In traditional process control, the solution is to use not a single control chart but to combine different control charts (such as Shewhart-CuSum [Luc82] and Shewhart-EWMA charts [LS90]) to detect shifts at different scales. This, of course, leads to increased alarm rates (false and true). The wavelet decomposition method offers a more elegant and suitable solution. Aradhye et al. [ABS03] used the term multiscale statistical process control (MSSPC) to describe these methods. Wavelet methods are also more suitable for autocorrelated data, since the wavelet decomposition can approximately decorrelate the measurements. A survey of wavelet-based process monitoring methods and their history is given in Ganesan et al. [GDV04]. Here we focus on their features that are relevant to biosurveillance.

The main application of wavelets has been for denoising, compressing, and analyzing image, audio, and video signals. Although wavelets have been used by statisticians for smoothing/denoising data (e.g., [DJ95], for density estimation [DJK96], nonparametric regression [OP96], and other goals [PW00]), they have only very recently been applied to statistical process monitoring. The most recent developments in wavelet-based monitoring methods have been published mainly within the area of chemical engineering [SCR97, HLM98, ABS03]. The main difference between chemical engineering processes and biosurveillance data (traditional and nontraditional) is that in the former the definitions of normal and abnormal are usually well-defined, whereas in the latter it is much harder to establish such clear definitions. In that sense wavelets are even more valuable in biosurveillance because of their nonspecific nature. Aradhye et al. [ABS03] have shown that using wavelets for process monitoring yields better average performance than single-scale methods if the shape and magnitude of the abnormality are unknown.

The typical wavelet monitoring scheme works in four main steps:

1. Decompose the series into coefficients at multiple scales using the discrete wavelet transform (DWT). The DWT algorithm is as follows:
   - Convolve the series with a low-pass filter to obtain the approximation coefficient vector $\mathbf{a}_1$ and with a high-pass filter to obtain the detail coefficient vector $\mathbf{d}_1$. If we denote the low-pass decomposition filter by $\mathbf{h} = [h_0, h_1, \ldots, h_n]$, then the ith component of the high-pass decomposition filter, $\mathbf{g}$, is given by $g_i = (-1)^i h_{n-i}$.
   - Downsample the coefficients. Half of the coefficients can be eliminated according to the Nyquist rule, since the series now has a highest frequency of $\pi/2$ radians instead of π radians. Discarding every other coefficient downsamples the series by two, and the series will then

have half the number of points. The scale of the series is now doubled [Pol].
- Reconstruct the approximation vector A_1 and detail vector D_1 by upsampling and applying "reconstruction" filters (Inverse-DWT). The set of low-pass and high-pass reconstruction filters are given as $h_n* = h_{-n}$ and $g_n* = g_{-n}$.

If an orthogonal wavelet is used, then the original signal can be completely reconstructed by simple addition: $Y = A_1 + D_1$. The second level of decomposition is obtained by applying this sequence of operations to the first level approximation A_1. The next levels of decomposition are similarly obtained from the previous level approximations.

2. Perform some operation on the detail coefficients $\mathbf{d}_k$ $(k = 1, \ldots, m)$. Various operations were suggested for monitoring purposes. Among them:
   - Thresholding the coefficients at each scale for the purpose of smoothing or data reduction [LLW02].
   - Forecasting each of the details and the mth approximation at time $t + 1$. This is done by fitting a model such as an autoregressive model [GSC02] or neural networks [AM97] to each scale and using it to obtain one-step-ahead forecasts.
   - Monitoring A_m and $D_1, D_2, \ldots, D_m$ by creating control limits at each scale [ABS03].
3. Reconstruct the series from the manipulated coefficients. After m levels of decomposition, an orthogonal wavelet will allow us to reconstruct the original series by simple addition of the approximation and detail vectors: $Y = A_m + D_1 + D_2 + \ldots + D_m$. If thresholding was applied, the reconstructed series will differ from the original series, usually resulting in a smoother series. In the case of single-scale monitoring [ABS03] use the control limits as thresholds and reconstruct the series only from the coefficients that exceeded the thresholds. In the forecasting scheme, the reconstruction is done to obtain a forecast of the series at time $t + 1$ by combining the forecasts at the different scales.
4. Perform some operation on the reconstructed series. Aradhye et al. [ABS03] monitor the reconstructed series using a control chart. In the forecasting scheme the reconstructed forecast is used to create an upper control limit for the incoming observation [GSC02].

Although DWT appears to be suitable for biosurveillance, it has several limitations that must be addressed. The first is that the downsampling causes a delay in detection and thus compromises timeliness. This occurs because the downsampling causes a lag in the computation of the wavelet coefficients, which increases geometrically as the scale increases. An alternative is to avoid the downsampling stage. This is called stationary- or redundant-DWT. Although it solves the delay problem, it introduces a different challenge; it does not allow the use of orthonormal wavelets, which approximately decorrelate the series. This means that we cannot treat the resulting coefficients at each

scale as normally distributed, uncorrelated, and with equal variance. Aradhye et al. [ABS03] conclude that for detecting large shifts it is preferable to use stationary-DWT if the series is uncorrelated or moderately correlated, whereas for highly nonstationary or autocorrelated series the use of downsampling is preferable. Both models perform similarly in detecting small changes. For further discussion of this issue and empirical results see Aradhye et al. [ABS03].

The second issue is related to the boundaries of the series, and especially the last observation. Since DWT involves convolving the series with filters, the beginning and end of the series need to be extrapolated (except when using the Haar). One approach is to use boundary-corrected wavelets. These have been shown to be computationally impractical [GDV04]. Another approach is to use an extrapolation method such as zero padding, periodic extension, and smooth padding. In surveillance applications the end of the series and the type of boundary correction are extremely important. Extrapolation methods such as zero padding and periodic extension (where the beginning and end of the series are concatenated) are clearly not suitable, since it is most likely that the next values will not be zeros or those from the beginning of the series. More suitable methods are the class of smooth padding, which consist of extrapolating the series by either replicating the last observation or linearly extrapolating from the last two values. An alternative would be to use exponential smoothing, which is known to have good forecasting performance in practice.

Finally, although wavelet-based methods require very little user input for the analysis, there are two selections that need to be made manually, namely, the depth of decomposition and the wavelet function. Ganesan et al. [GDV04] offer the following guidelines based on empirical evidence: the depth of decomposition should be half the maximal possible length. Regarding choice of wavelets, the main considerations are good time-frequency localizations, number of continuous derivatives (determine degrees of smoothness), and a large number of vanishing moments. We add to that computational complexity and interpretability. The Haar, which is the simplest and earliest wavelet function, is best suited for detecting step changes or piecewise constant signals. For detecting smoother changes, a Daubechies filter of higher order is more suitable.

4 Monitoring Multiple Data Streams

Modern biosurveillance systems such as the ones described earlier routinely collect data from multiple sources. Even within a single source there are usually multiple data streams. For instance, pharmacy sales might include sales of flu, allergy, and pain-relief medications, whereas ER visits record the daily number of several chief complaints. The idea behind syndromic surveillance is to monitor a collection of symptoms to learn about possible disease outbreaks. Therefore we expect multivariate monitoring methods to be superior

to univariate methods in actual detection, since the hypothesized signal can be formulated in a multivariate fashion. Optimally, multivariate models should detect changes not only in single data streams but also in the functional relationships between them.

4.1 Merging Data Sources: Why Use Aggregated Data?

One of the main reasons for treating biosurveillance data at the aggregated level is the issue of privacy associated with individuals whose data are being used. Medical and public health data systems of relevance for surveillance systems are typically subject to formal rules and/or legal restrictions regarding their use in identifiable form (e.g., as provided for by the Health Insurance Portability and Accountability Act of 1996, Public Law 104-191 (HIPAA) under its recently issued and implemented privacy and confidentiality rules), although there are typically research and other permitted uses of the data provided that they are de-identified. The HIPAA Safe Harbor de-identification, for instance, requires the de-identification of 18 identifiers including name, social security number, zip code, medical record number, age, etc. The removal of such information clearly restricts the possibility of record linkage across data sources, although it also limits the value of the data for statistical analysis and prediction, especially in connection with the use of spatial algorithms [Won04]. Similar legal restrictions apply to prescription information from pharmacies. Other public and semipublic data systems, such as school records, are typically subject to a different form of privacy restriction but with similar intent. Finally, grocery and OTC medication sales information is typically the property of the commercial interests that are wary of sharing data in individually identifiable form even if there are no legal strictures against such access. Solutions do exist that would potentially allow record linkage to at least some degree (e.g., by using a trusted broker and related data sharing agreements) (see the discussion in Gesteland et al. [GGT03]). While the practical solution of independently and simultaneously monitoring the separate sources, especially at the aggregate level, avoids the issue of record linkage and privacy concerns, it also leads to loss of power to detect the onset of a bioterrorist attack! Thus ultimately, if the syndromic surveillance methodology is to prove successful in early detection of a bioterrorist attack, the HIPAA and other privacy rules will need to be adjusted either to allow special exceptions for this type of data use, or to recognize explicitly that privacy rights may need to be compromised somewhat to better protect the public as a whole through the increased utility of the use of linked multiple data sources.

A separate reason for using aggregated data is the difficulty of record linkage from multiple sources: "identifiers" that are attached to records in different sources will usually differ at least somewhat. Linking data from two or more sources either requires unique identifiers that are used across systems or variables that can be used for record linkage. In the absence of unique

identifiers, matching names and fields, especially in the presence of substantial recording error, poses substantial statistical challenges. For further discussion of these issues see Fienberg and Shmueli [FS05] and especially Bilenko et al. [BMC03].

4.2 Current Common Approaches

Monitoring multiple time series is central in the fields of quality/process control, intrusion detection [Ye02], and anomaly detection in general. When the goal is to detect abnormalities in independent series, then multiple univariate tools can be used, and then merged to form a single alarm mechanism. However, the underlying assumption behind the data collected for biosurveillance is that the different sources are related and are measuring the same phenomenon. In this case, multivariate methods are more suitable. The idea is to detect not only abnormalities in single series, but also abnormal relationships between the series (also termed "counterrelationships"). In the following we describe multivariate methods that have been used in different applications for the purpose of detecting anomalies.

Statistical Process Control

The quality control literature includes several multivariate monitoring methods. Some are extensions of univariate methods, such as the χ^2 and Hotelling T^2 control charts, the multivariate CuSum chart, and the multivariate EWMA chart [ASJ97]. The multivariate versions are aimed at detecting shifts in single series as well as counterrelationships between the series. As in the univariate case, they are all based on the assumptions of independent and normally distributed observations. Also, like their univariate counterparts they suffer from problems of underdetection. In practice they are sometimes combined with a Shewhart chart, but this solution comes at the cost of slowing down the detection of small shifts [ASJ97]. When the multiple series are independent of each other, they do not require a multivariate model to monitor counterrelationships. An example is monitoring multiple levels of activity in an information system to detect intrusions, where Ye [Ye02] found that the different activity measures were not related to each other, and therefore a simple χ^2 chart outperformed a Hotelling T^2 chart. A multivariate model is still needed here, however, instead of a set of univariate control charts. One reason is the inflated false-alarm rate that results from multiple testing. If each of p univariate charts has a false-alarm probability α, then the combined false-alarm probability is given by

$$1 - (1 - \alpha)^p. \tag{13}$$

One solution is to use a small enough α in each univariate chart; however, this approach becomes extremely conservative and is impractical for the moderate to high number of series collected by biosurveillance systems. This issue is also

related to the problem of interpreting an alarm by the multivariate control chart. Although it may seem intuitive to determine which of the univariate measures is causing the alarm by examining the univariate charts, this is not a good approach not only because of the α-inflation but also because the alarm might be a result of changes in the covariance or correlation between the variables. Solutions for the α inflation based on Bonferroni-type adjustments have been shown to be conservative. A better approach is to decompose the monitoring statistic into components that reflect the contribution of each variable [Mon01]. For example, if the monitoring statistic is the Hotelling T^2, then for each variable i $(i = 1, \ldots, n)$ we compute

$$d_i = T^2 - T^2_{(i)}, \tag{14}$$

where $T^2_{(i)}$ is the value of the statistic for all the $p-1$ variables except the ith variable. This is another place where the use of FDR methodology may be appropriate and of help. One also needs to consider monitoring the covariance in parallel.

Other methods within this approach have tried to resolve the shortcomings of these control charts. One example is using Shewhart and CuSum charts to monitor "regression-adjusted variables," which is the vector of scaled residuals from regressing each variable on the remaining variables [Haw91]. Another example is a Bayesian approach for monitoring a multivariate mean (with known covariance matrix), where a normal prior is imposed on the process mean. A quadratic form that multiplies the posterior mean vector and the posterior covariance matrix is then used as the monitoring statistic [Jai93].

The second statistical approach towards multivariate monitoring is based on reducing the dimension of the data and then using univariate charts to monitor the reduced series and the residuals. PCA and partial least squares (PLS) are the most popular dimension reduction techniques. In PCA, principal components are linear combinations of the standardized p variables. We denote them by $PC_1, \ldots, PC_p$. They have two advantages. First, unlike the original variables the principal components are approximately uncorrelated. Second, in many cases a small number of components captures the variability in the entire set of data [NIS04]. Kaiser's rule of thumb for determining the number of components that is needed to capture most of the variability is to retain only components associated with an eigenvalue larger than 1 [Kai60]. There are alternatives, such as the number of components that explain a sufficient level of variability. In quality control usually the first two components, PC_1, PC_2, are plotted using a Hotelling-T^2 chart, but the number of components (k) can be larger. A second plot monitors the "residuals" using

$$Q = \sum_{i=k+1}^{p} \frac{PC_i^2}{\lambda_i}, \tag{15}$$

where λ_i is the eigenvalue corresponding to the ith principal component (which is also equal to its variance). Bakshi [Bak98] points out that these

charts suffer from the same problems of T^2 charts, as described above, in the sense of being insensitive to small changes in the process. Solutions are to monitor these statistics using a CuSum or EWMA scheme. The main shortcoming of these charts is their reliance on the assumption that the observations follow a multivariate normal distribution. In practice, multivariate normality is usually difficult to justify [CLL00]. This is especially true in biosurveillance where the different data sources come from very diverse environments.

Time Series Models

The multivariate form of ARMA models is called vector-ARMA models. The basic model is equivalent to (5), except that $\mathbf{y}_t$ $(t = 1, 2, \ldots)$ are now vectors. This structure allows for autocorrelation as well as cross-correlation between different series at different lags. In addition to the complications mentioned in relation to ordinary ARMA models, in vector-ARMA the number of α and θ parameters is larger ($(p + q + 1)$ multiplied by the number of series). The parameter covariance matrix to be estimated is therefore much larger. Since estimating the MA part adds a layer of complication, vector-AR models are more popular. In the context of biosurveillance, vector-AR models have advantages and disadvantages. Their strength lies in their ability to model lagged and counterrelationships between different series. This is especially useful for learning the pattern of delay between, for instance, medication sales and ER visits. However, vector-AR models have several weaknesses that are especially relevant in our context. First, their underlying assumption regarding the stationarity of the data is almost never true in data streams such as sales and hospital visits. This nonstationarity becomes even more acute as the frequency of the data increases. Second, although in some cases a set of transformations can be used to obtain stationarity, this preprocessing stage is highly series-specific and requires experience and special attention from the modeler. Furthermore, the application of different transformations can cause the series that were originally aligned to lose this feature. For example, by differentiating one series once and another series three times, the resulting series are of different length. Finally, any series that cannot be transformed properly into a stationary series must be dropped from the analysis. The third weakness of vector-AR models is that they are hard to automate. The model identification, estimation, and validation process is computationally heavy and relies on user expertise. Automated procedures do exist in software such as SAS (the VARMAX procedure [SAS00]). For determining the order of the model they use numerical measures such as Akaike information criterion (AIC), criterion final prediction error (FPE), and Bayesian information criterion (BIC). However, it is not guaranteed that the chosen model is indeed useful in a practical sense, and experienced statisticians would insist on examining other graphical measures such as auto- and cross-correlation plots to decide on the order of the model.

Estimation of the vector-AR parameters can be done using maximum likelihood. More often, for computational reasons, it is framed as an ordinary regression model and estimated using LS. Casting an AR model in the form of a regression model is an approximation in that in a regression model the predictors are assumed to be constant, whereas in an AR process they are a realization of a stochastic process. The parameter estimates are still consistent and asymptotically unbiased estimates for the AR model [NS01]. Thus, this estimation method is suitable for sufficiently long series, as is the case in typical biosurveillance data. However, collinearity and overparametrization are typical problems. One solution is to use a Bayesian approach and to impose priors on the AR coefficients [Ham94]. An alternative used by Bay et al. [BSU04] is to use ridge regression. The basic idea is to zero estimates that are below a certain threshold. Ridge regression yields biased estimates, but their variance is much smaller than their LS counterparts [MS75]. The estimated parameters are those that solve the equation

$$\boldsymbol{\beta} = (X'X + \lambda I)^{-1} X'\mathbf{y}, \tag{16}$$

where $\lambda \geq 0$ is the ridge parameter and I is the identity matrix. In the context of a vector-AR model we set $\mathbf{y} = \mathbf{y}_t$ (the multiple measurements at time t) and X is the matrix of lagged measurements at lags $1, \ldots, p$.

As with univariate ARIMA models, the stationarity assumption, the need in expert knowledge in model identification and estimation, the computational complexity, and overparametrization limit the usefulness of multivariate ARIMA models for integration into automated biosurveillance systems.

4.3 Alternative Approaches

Multichannel Singular Spectral Analysis

A generalization of SSA to multivariate time series, called multichannel-SSA (M-SSA), was described by Ghil and Yiou [GY96] and applied to several climate series. The idea is similar to the univariate SSA, except that now the lag-covariance matrix is a block-Toeplitz matrix T, where T_{ij} is an $M \times M$ lag-covariance matrix between series i and series j.

From a practical point of view, as the space increases in the number of series (L) and/or window width (M), the diagonalization of T, which is a $(T \times M) \times (T \times M)$ matrix, becomes cumbersome. Solutions include projecting the data onto a reduced subspace using PCA, undersampling the data to reduce M, and using expert knowledge to reduce the frequencies of interest. To give a feeling of the dimensions that can be handled, Plaut and Vautard [PV94] applied M-SSA to $L = 13$ series of 5-day observations, with $M = 40$ (equivalent to a maximum lag of 200 days).

There are several reasons why M-SSA should be investigated for biosurveillance. First, climatic data and syndromic data share components such as

weekly, seasonal, and annual patterns. Second, its relative insensitivity to the stationarity assumption makes it attractive for biosurveillance data. Finally, the ability to generalize it to the analysis of multiple time series (although computationally challenging) is useful not only for monitoring purposes but also for shedding light on the cross-relationship between different biosurveillance series, both within a data source and across data sources. The SSA-MTM toolkit is a software package for applying M-SSA (and other techniques), and is freely available at `http://www.atmos.ucla.edu/tcd/ssa/`.

Multivariate Wavelet Method

DWT has proven to be a powerful tool for monitoring nonstationary univariate time series for abnormalities of an unknown nature. Several authors created generalizations of the univariate method to a multivariate monitoring setting mostly by combining it with PCA. The most recent method, by Bakshi [Bak98], uses a combination of DWT and PCA to create a multiscale principal components analysis (MSPCA) for online monitoring of multivariate observations. The idea is to combine the ability of PCA to extract the cross-correlation between the different series with the wavelets' ability to decorrelate the autocorrelation within each series. As with control chart methodology, there are two phases: In phase I it is assumed that there are no abnormalities, and the control limits for the charts are computed. In phase II new data are monitored using these limits. The process used in MSPCA consists of

1. Decomposing each univariate series using DWT (the same orthonormal wavelet is used for all series).
2. Applying PCA to the vectors of coefficients in the same scale, independently of other scales.
3. Using T^2- and Q-charts to monitor the principal components at each scale. During phase I the control limits for these charts are computed.
4. Combining the scales that have coefficients exceeding the control limits to form a "reconstructed multivariate signal" and monitoring it using T^2- and Q-charts. During phase I the control limits for these two charts are computed. In phase II the reconstructed series is obtained by combining the scales that indicate an abnormality at the most recent time point.

The idea is that a change in one or more of the series will create a large coefficient first at the finest scale, and as it persists, it will appear at coarser scales (similar to the delay in detecting spike changes with CuSum and EWMA charts). This might cause a delay in detection, and therefore the reconstructed data are monitored in parallel. The overall control chart is used for raising an alarm, while the scale-specific charts can assist in extracting the features representing abnormal operation.

As in the univariate case, the downsampling operation causes delays in detection. Bakshi [Bak98] therefore suggests using a stationary-wavelet transform, which requires the adjustment of the control limits to account for the

coefficient autocorrelation that is now present and its effect on the global false-alarm rate. An enhancement to the Bonferroni-type adjustment suggested by Bakshi [Bak98] would be to use the more powerful FDR approach, which controls the expected proportion of false positives.

Multivariate Exponential Smoothing

Although research and application of univariate exponential smoothing is widespread there is a surprising scant number of papers on multivariate exponential smoothing, as a generalization of the univariate exponential smoothing methods. Two papers that have addressed this topic are Pfeffermann and Allon [PA89] and Harvey [Har86]. Since then, it appears, there has been little new on the topic.

The generalized exponential smoothing model suggested by Harvey [Har86] includes linear and polynomial trends and seasonal factors and can be estimated using algorithms designed for the univariate case. Pfeffermann and Allon [PA89] suggest a generalization of the Holt-Winters additive exponential smoothing, simply by expressing the decomposition and updating equations in matrix form. The only additional assumption is that the error term $\boldsymbol{\varepsilon}_t$ is assumed to have $\mathrm{E}(\boldsymbol{\varepsilon}_t) = \mathbf{0}$, $\mathrm{Var}(\boldsymbol{\varepsilon}_t) = \Sigma$, and $\mathrm{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon'}_{t-i}) = \mathbf{0}$ for $i > 0$. The set of updating equations are given by

$$\begin{aligned} \mathbf{L}_{t+1} &= A(\mathbf{Y}_{t+1} - \mathbf{S}_{t+1-M}) + (I - A)(\mathbf{L}_t + \mathbf{T}_t), \\ \mathbf{T}_{t+1} &= B(\mathbf{L}_{t+1} - \mathbf{L}_t) + (I - B)\mathbf{T}_t, \\ \mathbf{S}_{t+1} &= C(\mathbf{Y}_{t+1} - \mathbf{L}_{t+1}) + (I - C)\mathbf{S}_{t+1-M}, \end{aligned} \tag{17}$$

where A, B, and C are three convergent matrices of smoothing constants. The m-step-ahead forecast at time t is

$$\hat{\mathbf{Y}}_{t+m} = \mathbf{L}_t + m\mathbf{T}_t + \mathbf{S}_{t+m-M}. \tag{18}$$

These are similar to the univariate smoothing updating and prediction equations. In fact, the updating equations can be written as weighted averages of estimates derived by the univariate components and correction factors based on information from the other series (the off-diagonal elements of the matrices A, B, and C). Pfeffermann and Allon [PA89] show that the forecasts from this model are optimal under particular state space models. They also illustrate and evaluate the method by applying it to two bivariate time series: one related to tourism in Israel, and the other on retail sales and private consumption in Israel. They conclude that the multivariate exponential smoothing (MES) forecasts and seasonal estimates are superior to univariate exponential smoothing and comparable to ARIMA models for short-term forecasts. Although the model formulation is distribution free, to forecast all series the specification of the smoothing matrices and initial values for the different components requires a distributional assumption or prior subjective

judgments (which are much harder in a multivariate setting). This is the most challenging part of the method. However, once specified, this process need not be repeated. Also, once specified, the estimated smoothing matrices can shed light on the cross-relationships between the different time series in terms of seasonal, trend, and level components.

Data Depth

The pattern recognition literature discusses nonparametric multivariate models such as those associated with data depth methodology. This approach was developed through techniques at the interface between computational geometry and statistics and is suitable for nonelliptically structured multivariate data [Liu03]. A data depth is a measure of how deep or how central a given point is with respect to a multivariate distribution. The data depth concept leads to new nonparametric, distribution-free multivariate statistical analyses [RS04a], and in particular, it has been used to create multivariate monitoring charts [Liu03, LS02]. These charts allow the detection of both a location change and a scale increase in the process simultaneously. In practice, they have been shown to be more sensitive to abnormalities relative to a Hotelling-T^2 chart in monitoring aviation safety, where the data are not multivariate normal [CLL00]. There are several control charts that are based on data depth measures, the simplest being the r chart. In this time-preserving chart the monitoring statistic is the rank of the data depth measure, denoted by r. Liu and Singh [LS93] proved that r converges in distribution to a U(0,1) distribution. Therefore, the lower control limit on the r-chart equals the α of choice, and if the statistic exceeds this limit, it means that the multivariate observation is very far from the distribution center, and a flag is raised. The computation of the data depth measures becomes prohibitively intensive as the dimension of the space increases. Solutions have been to use probabilistic algorithms [CC03].

4.4 Spatial Approaches to Biosurveillance

A different approach to monitoring multiple data sources has been to focus on the spatial information and look for geographical areas with abnormal counts. Two major approaches have been used for monitoring biosurveillance data using a spatial approach. Both operate on discrete, multidimensional temporal datasets. The first method uses the algorithm What's Strange About Recent Events (WSARE), which is applied in RODS and uses a rule-based technique that compares recent emergency department admission data against data from a baseline distribution and finds subgroups of patients whose proportions have changed the most in the recent data [WMC03]. In particular, recent data are defined as all patient records from the past 24 hours. The definition of the baseline was originally the events occurring exactly five, six, seven, and eight weeks prior to the day under consideration (WSARE version 2.0) [WMC02]. Such

a comparison eliminates short-term effects such as day-of-week, and longer-term seasonal effects (by ignoring weeks that are farther in the past). The baseline was then modified to include all historic days with similar attributes (WSARE version 2.5), and in the current version (WSARE 3.0) a Bayesian Network represents the joint distribution of the baseline [WMC03]. One limitation of WSARE is that it is practically limited to treating a maximum of two rules (i.e., variables), due to computational complexity [WMC02, WMC03]. Another computational limitation is the randomization tests used to account for the multiple testing, which are also computationally intensive. Finally, WSARE can use only discrete data as input, so that continuous information such as age must be categorized into groups. This, of course, requires expert knowledge and might be specific to the type of data monitored and/or the outbreak of interest.

A different method, implemented in ESSENCE II and in the NYC-DOHMH system, is the spatial-temporal scan statistic [Kul01], which compares the counts of occurrences at time t in a certain geographical location with its neighboring locations and past times, and flags when the actual counts differ consistently from the expected number under a nonhomogeneous Poisson model. The purely spatial approach is based on representing a geographical map by a uniform two-dimensional grid and aggregating the records within families of circles of varying radii centered at different grid points. The underlying assumption is that the number of records in a circle come from a nonhomogeneous Poisson process with mean qp_{ij} where q is the underlying disease rate and p_{ij} is the baseline rate for that circle. The purely spatial scan statistic is the maximum likelihood ratio over all possible circles, thereby identifying the circle that constitutes the most likely cluster. This requires the estimation of the expected number of cases within each circle and outside of it given that there is no outbreak. The p-value for the statistic is obtained through Monte Carlo hypothesis testing [Kul01]. The spatial-temporal scan statistic adds time as another dimension, thereby forming cylinders instead of circles. The varying heights of the cylinders represent different windows in time. The multiple testing is then accounted for both in space and in time domains. Lawson [Law01] mentions two main challenges of the spatial-temporal scan statistic, which are relevant to biosurveillance. First, the use of circular forms limits the types of clusters that can be detected efficiently. Second, the timeliness of detection and false-alarm rates need further improvement. In an application of the scan statistic to multiple data sources in ESSENCE II, Burkom [Bur03b] describes a modified scan-statistic methodology where the counts from various series are aggregated and used as the monitored data, and these are assumed to follow an ordinary Poisson model. A few modifications were needed to address features of biosurveillance data. The uniform spatial incidence is usually inappropriate and requires the estimation of expected counts for each of the data sources (which is challenging in and of itself); the aggregation of counts from disparate sources with different scales was adjusted by using a "stratified" version of the scan statistic. It appears

that such data-specific and time-varying tailoring is necessary and therefore challenges the automation of this method for biosurveillance.

Both methods are flexible in the sense that they can be applied to different levels of geographical and temporal aggregation and for different types of diseases. With respect to automation and user input the two methods slightly differ. In the scan-statistic methods the user is required to specify the maximal spatial cluster size (represented by the circle radius) and the maximal temporal cluster length (represented by the cylinder height). In addition, since neither the Poisson nor the Bernoulli model is likely to be a good approximation of the baseline counts in each area, a nonhomogeneous Poisson will most likely be needed. This requires the specification of the relevant variables and the estimation of the corresponding expected counts inside and outside each cylinder. For WSARE the user need only specify the time window that is used for updating the Bayesian network. Finally, the major challenge in these two spatial methods as well as other methods (e.g., the modified Poisson CuSum method by Rogerson [Rog01]) is their limitation to monitoring only count data and the use of just categorical covariates.

5 Concluding Remarks

The collection of data streams that are now routinely collected by biosurveillance systems is diverse in its sources and structure. Since some data sources comprise multiple data streams (e.g., different medication sales or different chief complaints at ER admission), there are two types of multivariate relationships to consider: "within sets" and "across sets." Data streams within a single source tend to be qualitatively more similar to each other as they are measured, collected, and stored by the same system and share common influencing factors such as seasonal effects. Data streams across different sources are obviously less similar, even if the technical issues such as frequency of measurement and missing observations are ignored. The additional challenge is that the signature of a disease or bioterrorism-related outbreak is usually not specified and can only be hypothesized for some of the data sources (e.g., how does a large-scale anthrax attack manifest itself in grocery sales?). Stoto et al. [SFJ06] discuss the utility of univariate methods in biosurveillance by comparing univariate and multivariate Shewhart and CuSum chart performance. Their discussion and analyses are provocative, but there is need for a serious testbed of data to examine the utility of the different approaches.

The task of monitoring multivariate time series is complicated even if we consider a single data source. Traditional statistical approaches are based on a range of assumptions that are typically violated in syndromic data. These range from multivariate normal distribution, independence over time, to stationarity. Highly advanced methods that relax these assumptions tend to be very complicated, computationally intensive, and require expert knowledge to apply them to real data. On the other hand, advances in other fields where au-

http://www.springer.com/978-0-387-32904-8

Statistical Methods in Counterterrorism
Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication
Wilson, A.; Wilson, G.; Olwell, D.H. (Eds.)
2006, XII, 292 p. 14 illus., Softcover
ISBN: 978-0-387-32904-8