

Protecting against Biological Terrorism: Statistical Issues in Electronic Biosurveillance

Ronald D. Fricker, Jr., and Henry Rolka

In the post-9/11 world, biological terrorism (or bioterrorism) is of increasing concern to government and public health officials. To address this threat adequately, policymakers have recognized it is important to put in place systems that provide the earliest possible detection of an act of bioterrorism and real-time, ongoing information about its location and spread so the public health and medical communities can efficiently and effectively treat those affected.

Focusing on just the goal of early detection, it is a difficult problem to reliably extract a bioterrorism event “signal” from spatially and temporally varying natural disease incidence “noise” while simultaneously maintaining a low false signal rate. For example, in “Syndromic Surveillance: Is It Worth the Effort?”—published in *CHANCE*, vol. 17—Stoto, Schonlau, and Mariano evaluated the performance of a number of statistical methods to temporally detect a simulated bioterrorism event and concluded, “Our simulation study gives some indication of the size and speed that outbreaks must attain before they are detectable, and the results are sobering.”



What Is Electronic Biosurveillance?

Syndromic surveillance is defined by the Centers for Disease Control and Prevention (CDC) as “...surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response.” While this is not a universally accepted definition—it is, in fact, considered a bit outdated by some at the CDC—it is a worthwhile place to start our discussion, as it focuses on a number of important ideas.

First, syndromic surveillance is health surveillance, not military, regulatory or intelligence surveillance. It uses “health-related” data, a construct that encompasses a variety of data, including counts of individuals coming into medical facilities, over-the-counter medication sales, and aggregate laboratory test results. Second, the data and associated surveillance generally precede diagnosis or ‘case’ confirmation. The goal of a syndromic surveillance system is to gather data in advance of diagnostic case confirmation to give early warning of a possible outbreak. Clearly, once a definitive diagnosis has been made, the need for detection becomes moot. Third, the process must provide a signal of “sufficient probability” to trigger “further public health response.” Syndromic surveillance is not supposed to provide a definitive determination that an outbreak is occurring. Rather, it is supposed to signal that an outbreak may be occurring, indicating a need for further evidence or triggering an investigation by public health officials (i.e., the CDC or a local or state public health department).

Broadly speaking, syndromic surveillance is intended to provide early disease or bioterrorism warning. However, the term syndromic surveillance as currently used is something of a catch-all for various types of new surveillance systems and paradigms that are now feasible due to mainly advances in computer and information technology. Biosurveillance has a much longer history for naturally occurring diseases (as opposed to the deliberate, malicious release of a toxin) and other traditional public health phenomena. Indeed, the field of epidemiology traces its roots back to John Snow, a London physician in the 1800s. Dr.



Figure 1. John Snow’s map. Pump locations are indicated by circled dots, and the bars represent the number of cholera deaths at each location. In the center of the map (and in the center of a large cluster of deaths) is the Broad Street pump.

Snow determined cholera was transmitted by a contaminated water supply in the London cholera epidemic of 1854. One way he demonstrated this was to simultaneously plot on a map the number of cholera deaths by city address and the locations of each of the city’s water pumps. As shown in Figure 1, the result was a clear visual association between areas of higher death rates from cholera and certain water pumps.

In many ways, we can think of syndromic surveillance as the modern, automated equivalent of Dr. Snow’s map. Two important differences, which we will return to later in this article, are that today’s syndromic surveillance systems are being put in place in advance of an outbreak and they are intended to actively search for possible outbreaks. Traditionally, epidemiological studies—including Dr. Snow’s analysis—are retrospective studies initiated to address a specific population risk event. Syndromic surveillance is intended to be prospective. Conducting a retrospective study to understand the cause of a disease or outbreak is hard enough; trying

to define a system that will effectively detect a variety of possible outbreaks that have not yet occurred—from naturally occurring diseases to bioterrorism attacks—is a daunting task.

In the post-9/11 world, many believe there is a critical need for syndromic surveillance. The CDC states on its syndromic surveillance web site, “Though, historically, syndromic surveillance has been utilized to target investigation of potential cases, its utility for detecting outbreaks associated with bioterrorism is increasingly being explored by public health officials.” More recently, syndromic surveillance and bioterrorism outbreak detection are considered part of the broader concept of electronic biosurveillance that encompasses both “early event detection” and “situational awareness.” Early event detection includes the detection of both natural and man-made health events using syndromic surveillance as well as many other types of information and data. Situational awareness is the real-time analysis and display of health data to monitor the location, magnitude, and spread of an

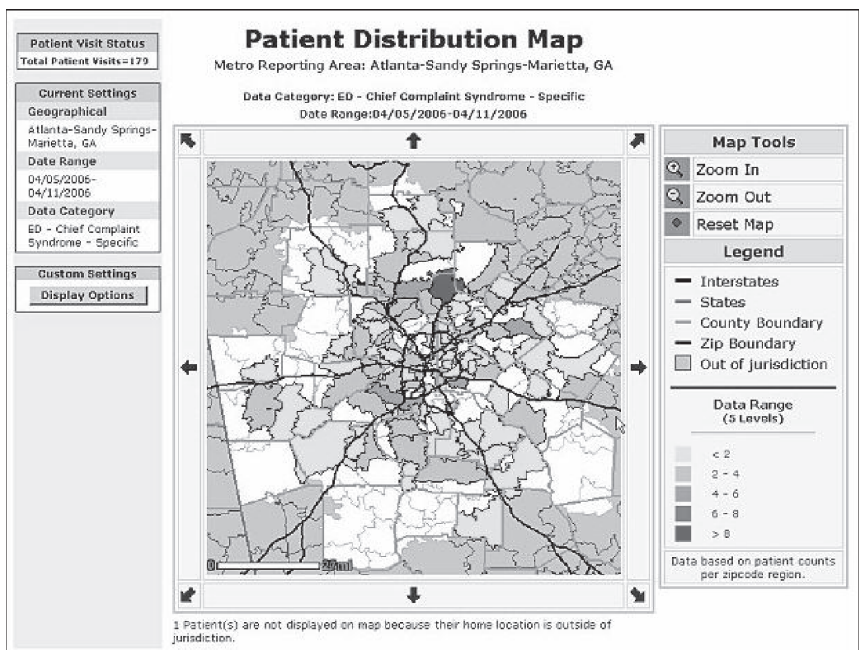
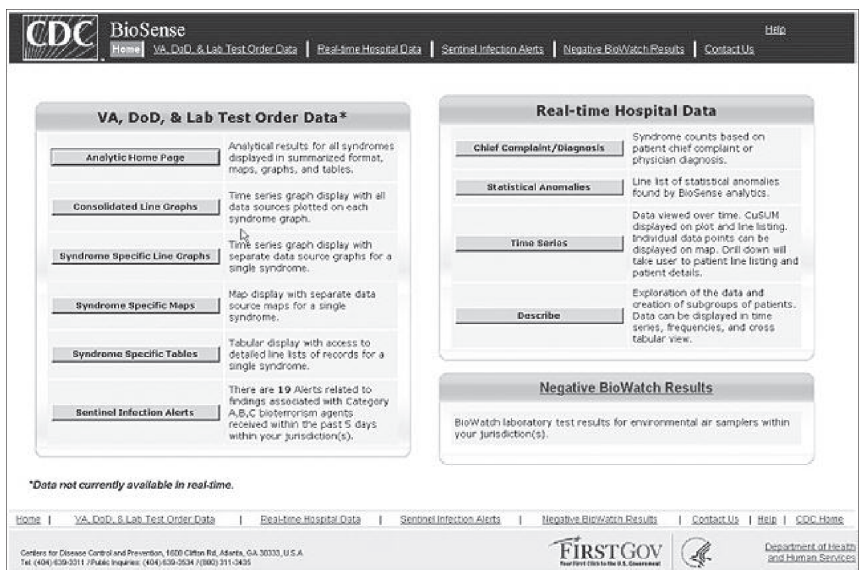


Figure 2. Screen shots of CDC’s BioSense system, taken from www.cdc.gov/biosense and www.cdc.gov/biosense/publichealth.htm

outbreak, in addition to the availability and application of public health and medical resources in response to the outbreak. Succinctly put, early event detection is about finding and identifying an outbreak, and situational awareness is about managing the response to it. The CDC and many state and local health departments are actively developing and fielding electronic biosurveillance systems, such as the CDC’s BioSense system shown in Figure 2. As described on the CDC’s BioSense web site (www.cdc.gov/biosense):

“BioSense is a national program intended to improve the nation’s capabilities for conducting near real-time biosurveillance, enabling health situational awareness through access to existing data from health care organizations across the country. The primary objective is to expedite event recognition and response coordination among federal, state, and local public health and health care organizations by providing each level of public health access

to the same data, at the same time. In other words, if a bioterrorism event or a disease outbreak occurs, every level of public health will be able to see health care data from their community in near real-time.”

However, many challenges in developing and fielding electronic biosurveillance systems, particularly large-scale systems, remain. Some of these challenges involve legal and regulatory issues, some involve personal privacy and proprietary information issues, and others are related to the technological and managerial challenges of assembling and managing such systems. And there are also a plethora of statistical challenges that must be solved to field and operate effective electronic biosurveillance systems.

An Example: Applying Existing Statistical Methods

In 1931, in *Economic Control of Quality of Manufactured Product*, Walter A. Shewhart developed the concept of the control chart, a graphical statistical tool most commonly used to control manufacturing processes. The success of Shewhart’s method lies in its simplicity. Essentially, one establishes control limits and, so long as various statistics derived from the data—sequentially observed over time—fall within the control limits, the process is assumed to be in control. If one or more fall outside the control limits (or perhaps in certain patterns), then the process is examined to determine whether it is out of control and requires adjustment. Shewhart’s work gave rise to the field of statistical process control (SPC) and a large and still growing literature of research into myriad statistical methods for controlling processes.

Shewhart’s method often is thought of as a sequential hypothesis test. At each time period, a statistical test is conducted to determine whether the underlying population parameter—often the mean—has changed. As any hypothesis test (and all SPC methods), it is subject to Type I and Type II errors. In the context of SPC, a Type I error occurs when the methodology indicates the process is out of control when, in fact, it is not. A Type II error occurs when the method should signal

the process is out of control, but fails to do so. And, because in SPC we are conducting a continuing sequence of tests until a signal is raised, a Type I error (often referred to as a false alarm) must eventually occur—even if the process always stays in control.

In the SPC literature, the metric most commonly used to evaluate the performance of an SPC method is the average run length, or ARL. The ARL under the assumption that the parameter of interest has not changed is referred to as the in-control ARL; it is the average time to a false alarm. When comparing SPC methods, the control limits are set first so the various methods have equal in-control ARLs, then their out-of-control ARLs for various out-of-control conditions are compared. For a given in-control ARL, methods with shorter out-of-control ARLs are preferred.

Advances in Statistical Process Control

There have been many advances in SPC since Shewhart's seminal work. Indeed, over decades of research, many methods have been developed and evaluated, particularly in the context of controlling manufacturing processes. In fact, the various quality-control movements (e.g., TQM) of the past and the ongoing Six Sigma methodology have their roots in statistical process control.

Two popular alternatives to Shewhart's methods are the cumulative sum (or CUSUM) and the exponentially weighted moving average (EWMA). Roughly speaking, the CUSUM works by summing up observations over time (though there are important details to the method that involve subtracting a quantity at each time period and "reflecting" the cumulative sum at zero). The EWMA is also a summation of the data over time, but of a different sort. At each time period, the EWMA calculates a weighted average of the EWMA's value from the previous time period and current observation.

For our purposes, the specific mathematical details of the methods are not as important as the fact that each procedure has different strengths. Shewhart's method generally outperforms the others when the out-of-control shift is large. In contrast, the CUSUM is usually better when the out-of-control shift is small

and sustained. And the EWMA can be made to perform much like either the Shewhart or the CUSUM, depending on how a weighting parameter is set.

Most of the well-known methods are univariate in nature, meaning they operate on only one stream of data—say the output from one manufacturing process or from one machine. However, multivariate methods also have been developed that operate on multiple data streams simultaneously. For example, in 1947, Harold Hotelling published a Shewhart-like methodology for multivariate observations that was an outgrowth of research related to the manufacture of bomb sights in World War II. Since that time, research has included the development of various multivariate generalizations of the CUSUM and EWMA, as well as other multivariate SPC methods.

Introduction to Statistical Quality Control, by Doug Montgomery, is the authoritative introduction to the field of SPC. "The Use of Control Charts in Health Care and Public Health Surveillance" by Bill Woodall that recently appeared in the *Journal of Quality Technology* is an excellent review of the use of control charts broadly applied to health care and public health surveillance.

SPC Applied to Syndromic Surveillance

An informed rational decisionmaking strategy is needed to identify and respond to analytic health risk indications in biosurveillance data. One component of this strategy is to apply existing statistical process control methods to health-related data. This approach is not without difficulty and controversy, because the health-related data and the underlying "disease processes" are generally more complicated and less controllable than manufacturing processes for which SPC was developed and to which it has been traditionally applied. However, the standard SPC methods are being applied in actual syndromic surveillance systems, and so here we illustrate their application and, in particular, how the performance of these methods is evaluated in the context of syndromic surveillance.

Specifically, we consider the application of simultaneous univariate

CUSUMs and a multivariate CUSUM (MCUSUM) method designed to look for increases of disease incidence in emergency room "chief complaint" data. (For readers interested in the technical details of the methods, and for additional results, see "Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance" in *Advances in Disease Surveillance* at www.isdsjournal.org.) A chief complaint is a broad categorization that captures the primary symptom or reason the patient sought care, such as respiratory complaint, gastrointestinal complaint, or unspecified infection. The purpose of the comparison is to evaluate whether it would be more effective to use simultaneous individual CUSUMs, each applied to a different data stream—say each type of chief complaint at each hospital—or whether a multivariate method that evaluates a series of data streams—say all chief complaints at each hospital or one type of chief complaint across multiple hospitals—performs better. In terms of the evaluation, "better" means that for a given in-control ARL fixed for all procedures being compared, the procedure that tends to have the shortest out-of-control ARLs across a range of out-of-control conditions would be the better method. This notion also may be roughly thought of as the best sensitivity over a range of conditions for a set level of specificity.

Figure 3 shows how the two methods—the multivariate CUSUM ("mod. MCUSUM") and the simultaneous univariate CUSUMs ("Indiv. CUSUMs")—can be compared via simulation. Here, observations were simulated from a six-dimensional multivariate standard normal distribution in Mathematica, where the random observations were generated using the "MultinormalDistribution" function. In this comparison, the out-of-control condition consisted of the mean increasing linearly in one dimension by δ (as shown on the x-axis) at each time period, with the in-control ARLs (where for the in-control mean we have $\delta = 0$) set to 100.

What this simulation result shows is, for the out-of-control condition specified, the simultaneous univariate CUSUMs outperform the multivariate CUSUM for all possible δ s. Other simu-

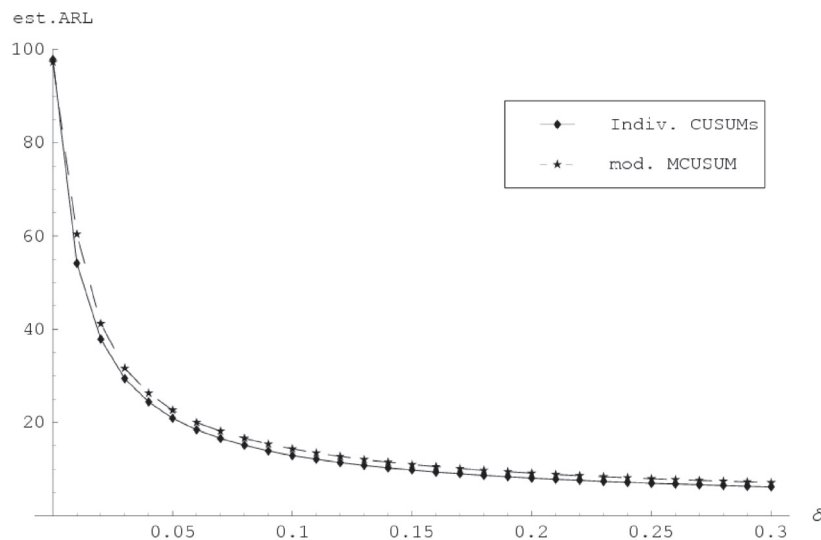


Figure 3. Results of a simulation comparison of a multivariate CUSUM (mod. MCUSUM) method versus the application of simultaneous univariate CUSUMs (Indiv. CUSUMs), one to each dimension. The out-of-control condition consisted of the mean in increasing in one dimension linearly by δ at each time period with the in-control ARLs set equally to 100. For this out-of-control condition, the simultaneous univariate CUSUMs had smaller ARLs than the multivariate CUSUM for all possible δ s.

lations for different types of changes in the mean, for example a linear increase in multiple dimensions, result in the MCUSUM outperforming the simultaneous individual CUSUMs.

The point here is that conducting a series of these types of simulations, which explore a variety of out-of-control behaviors, provides insight into the performance characteristics of the various methods. That is, under the idealized simulation conditions, it provides researchers and practitioners with information about whether and when each method should be preferred. This type of simulation comparison is commonly conducted in the industrial SPC literature. However, they are not widely accepted in the public health world for a number of reasons, including a lack of information about how to realistically simulate the “in-control” behavior of health data.

To address this criticism, one might also demonstrate how the procedures perform under real-world conditions by applying the methods to actual data—in this case, from five hospitals located in a large metropolitan area. The data consists of respiratory chief complaint counts by hospital for 2.5 years from October 1, 2001, to March 31, 2004.

Applying the procedures to these data, which capture naturally occurring incident rates and variation within hospitals and the covariation between hospitals, provides insight into real-world performance (although it, too, is subject to various challenges).

The Data

Figure 4 shows estimated mean respiratory chief complaint counts by hospital. Each point on each line is a four-week moving average using the data for two weeks prior and two weeks after the date plotted. A number of features of the data are clear from the figure, including:

- The hospital moving averages do not exhibit an increasing or decreasing trend, indicating the long-term incidence rate for respiratory chief complaints is relatively constant.
- Yet, there are visible ‘events’ in the data that persist for periods of time. For example, there are peaks across most or all the hospitals in January–February 2002 and December 2003–January 2004 that likely correspond to flu outbreaks.
- These events are consistent with the CDC’s aggregate data on “percent-

age of visits for influenza-like illness reported by sentinel physicians” for the South Atlantic region of the United States, where large outbreaks occurred in February–March 2001 and December 2003–January 2004, with a much smaller flu season outbreak in mid-February to mid-March 2003.

- The hospital counts are weakly positively correlated. Using the first six months of the data, the covariances between all pairs of hospitals were between 0.0 and 0.24.

In addition, there are significant differences in mean counts between hospitals, indicating some hospitals either serve larger populations or serve populations with higher respiratory illness rates (or both), as well as significant variation in the raw counts (see Figure 5) around the smoothed mean.

Evaluating Performance

Figure 5 displays the signal times for the two methods—the MCUSUM and the simultaneous univariate CUSUMs—when they are run retrospectively on the respiratory data from April 1, 2002, to March 31, 2004. The figure shows the raw data with the smoothed means (of Figure 4) and first signal times overlaid. (“First signal time” means repeated signals within 30 days of the first signal are suppressed for plot clarity.) The signal times for the MCUSUM are indicated by the dark vertical lines with the specific dates at the top. The signal times for the simultaneous individual CUSUMs are indicated by the diamonds plotted on the relevant smoothed mean. For example, Figure 5 shows the first signal for the MCUSUM occurred on November 15, 2002, and one of the individual CUSUMs also signaled on the same day.

What Figure 5 generally shows is that the MCUSUM and the simultaneous individual CUSUMs performed similarly on this data. As we just discussed, both schemes signaled on November 15, 2002. Similarly, on December 16, 2003, the MCUSUM signaled, after which four of the individual CUSUMs signaled, the earliest of which was also on December 16. And, on September 26, 2003, the MCUSUM signaled, after

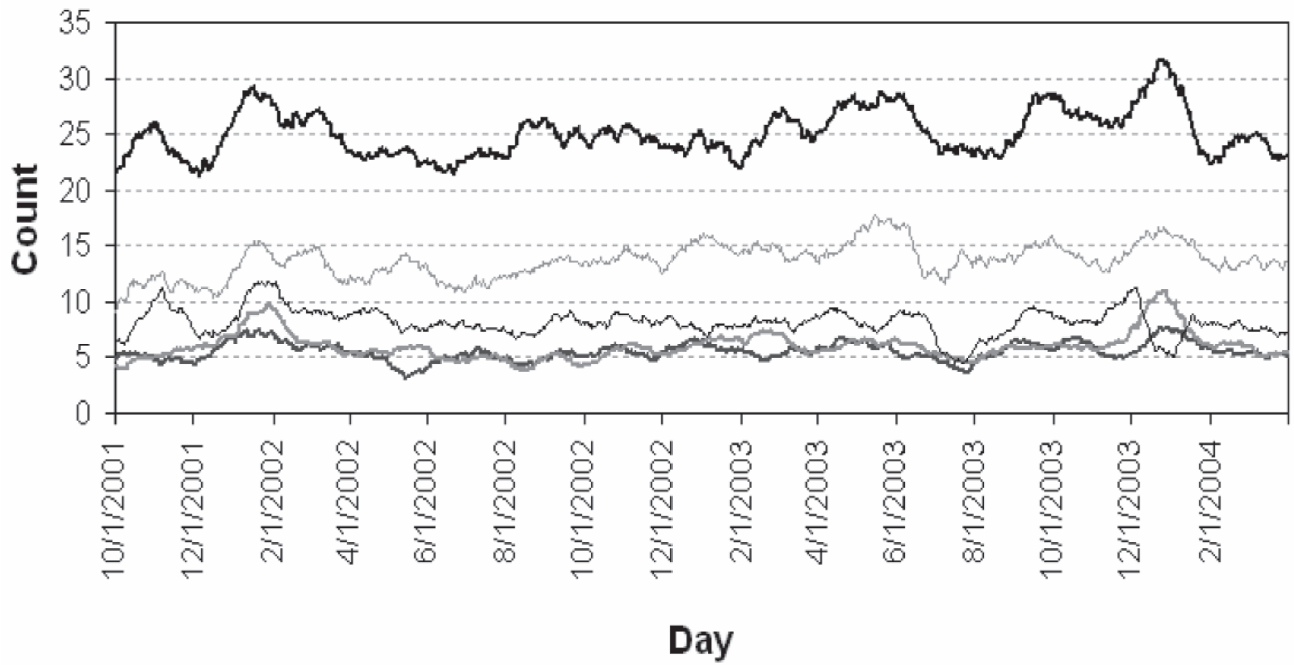


Figure 4. Smoothed respiratory “chief complaint” counts by hospital. Each point is a four-week moving average using the data for two weeks prior and two weeks after the date, and each line represents one hospital.

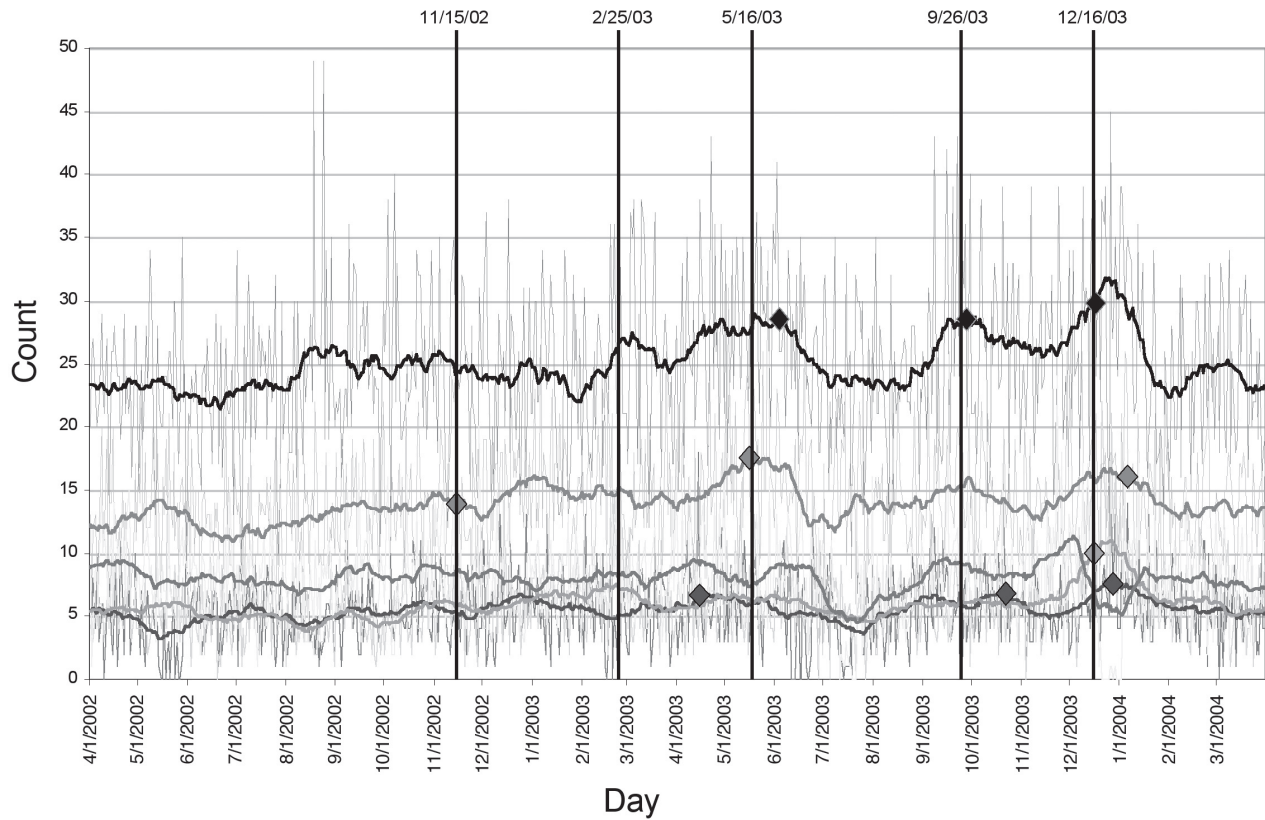


Figure 5. Plot showing when the MCUSUM and individual CUSUMs first signaled when run on the actual respiratory chief complaint data. The vertical lines are the signal times for the MCUSUM. The diamonds indicate the signal times for each CUSUM. “First signaled” means repeated signals within 30 days of the first signal are suppressed for plot clarity.

which two of the individual CUSUMs signaled, the earliest at September 27—a difference of just one day.

There are two MCUSUM signals that diverge from the simultaneous individual CUSUMs. On February 25, 2003, the MCUSUM signaled, while none of the individual CUSUMs did. And, on May 16, 2003, the MCUSUM signaled, before which two individual CUSUMs signaled (April 14 and May 15), but also after which one CUSUM signaled, on May 27. In spite of these small differences, the general result is that both the MCUSUM and the simultaneous individual CUSUMs perform very similarly with consistent signals.

While the multivariate CUSUM and simultaneous individual CUSUMs exhibited similar performance on the real hospital data, in more abstract simulation evaluations (such as in Figure 4), each seemed to demonstrate a separate specific strength: The modified MCUSUM is slightly better at detecting small shifts in many or all dimensions, while the simultaneous individual CUSUMs seem better at detecting a shift in only one dimension. This suggests a strategy of using both in combination, where in the public health arena, for example, individual hospitals might monitor their own trends using individual CUSUMs, while a city, county, or state public health department might monitor an area using the MCUSUM. Or, perhaps a public health department might use both the individual CUSUMs and the MCUSUM simultaneously, but interpret their signals differently: An individual CUSUM signal indicates the possibility of a localized event, while a MCUSUM signal indicates the possibility of a larger, area-wide event.

Many (Statistical) Issues Remain To Be Solved

While, at first glance, the application of SPC methods to health-related data seems to be an immediate solution to the biosurveillance problem, electronic biosurveillance is a less well-defined and more operationally complex problem than those for which SPC was originally developed, hence there are many technical and other issues that need to be solved or resolved. One way to group the challenges is to consider three broad

categories: definitional, developmental, and implementational. Important quantitative and statistical issues cut across all these areas, but particularly the first two.

Definitional Challenges

As we mentioned in the introductory section, the term syndromic surveillance is something of a catch-all for various types of new health surveillance. Such systems were originally intended to detect abnormalities in health-related data that would warrant further evidence or public health investigation. As first conceived, the procedures were expected to be capable of reliably detecting abnormalities representing medium to large population events in a relatively short time, sometimes expressed in hours or a small number of days. However, some research has raised questions that are driving syndromic surveillance systems development to optimize outbreak detection and response in conjunction with more traditional approaches, such as the 'sentinel physician.' (For example, an 'influenza sentinel physician' provides regular reports to public health agencies during the flu season on the number of patient visits each week and the number of patients with influenza-like illness. For syndromic surveillance, one can imagine a similar system of alert clinicians that report unusual findings directly to the public health communities.)

At issue is whether syndromic surveillance systems as originally conceived can achieve the necessary speed of outbreak detection while simultaneously maintaining reasonably low false alarm rates. This is partially due to the fact that electronic biosurveillance, as compared to other surveillance paradigms, lacks the maturity that comes with operational refinement over decades of use. It is also due to the need for progress in advancing the technological enablers for data acquisition timeliness and geographic coverage. And, as we have previously discussed, there is a need for additional methodological development in order to fully exploit the information embedded in these data and to statistically characterize the degree of event (un)certainly in order to determine when to engage a public health response to a signal.

When Does an Electronic Biosurveillance System Add Value?

Speed-of-detection questions aside, a critical question statisticians can help the public health community address is under what circumstances the various statistical detection methods can usefully detect an outbreak at all. Figure 6 illustrates the issue. Remember that one goal of biosurveillance is early event detection in advance of a medical diagnosis. Hence, if the outbreak is sufficiently large, geographically concentrated, and/or easy to diagnose, a medical or public health practitioner is likely to be equally fast or faster at detecting the outbreak than a statistical algorithm. In contrast, if the outbreak is very small and/or diffuse, a statistical algorithm operated in isolation is unlikely to detect the outbreak—at least in the short time desired of a syndromic surveillance system. The result of these restrictions is that statistical methods will be of value only when the outbreak is large/concentrated enough to statistically detect, but not so large that the outbreak is obvious, combined with the situation where identification of the type of outbreak is sufficiently hard to diagnose, making the medical profession likely to miss it for some time.

Thus, in conjunction with the public health community, statisticians can contribute to the development of electronic biosurveillance, syndromic surveillance, and other early event detection systems by helping to determine under what conditions and for what types of outbreaks the dotted region in Figure 6 exists. Statisticians also can help determine when it is of sufficient size that the implementation of a particular detection method for a particular type of outbreak in a biosurveillance system is worthwhile. This is a clear role and responsibility for statisticians and the statistical community in advancing biosurveillance effectiveness in efforts to protect the public health.

What Else Might an Electronic Biosurveillance System Do?

Whether or not syndromic surveillance or other early event detection systems turn out to be useful for outbreak detection, appropriately designed systems also may be useful for establishing and maintaining situational aware-

ness. Situational awareness occurs when an individual or organization has a complete understanding of his/her/its environment. It is particularly important for effective decisionmaking in complex, dynamic environments that are constantly changing. In terms of public health surveillance, in the event of an outbreak, public health officials will require critical information about the location, magnitude, and, in the case of a communicable agent, spread of the agent to effectively mitigate the effects of the outbreak. Only with superior situational awareness will public health officials be able to quickly and effectively manage their resources, identify outbreak sources and causes, identify affected and at-risk populations, and effectively isolate contagious populations.

However, note that it does not follow that electronic biosurveillance systems designed and optimized for early event detection will necessarily be effective at providing situational awareness or vice versa. For example, effective early event detection is more analytically focused, requiring sophisticated statistical methods to signal unusual trends and clusters and a system for public health officials to easily dig into the data once a signal has been raised in order to quickly determine whether an actual outbreak has occurred. In contrast, systems designed to promote situational awareness are likely to be more focused on descriptive statistics, and will probably require sophisticated data and information displays that effectively summarize the status of public health resources and population health indicators.

These possible foci for an electronic biosurveillance system, while related, are not the same. A common feature requirement for both purposes is the ability to quickly 'slice and dice' through data, creating views that support decisions about a variety of query types. Statisticians can help the public health community clearly distinguish between these functions, their requirements, and the necessary features of a system designed to support one or the other of these goals. A key to fulfilling this responsibility and providing such a service is the ability to communicate and interact successfully with stakeholders who may not be trained in statistical thinking but who are domain experts

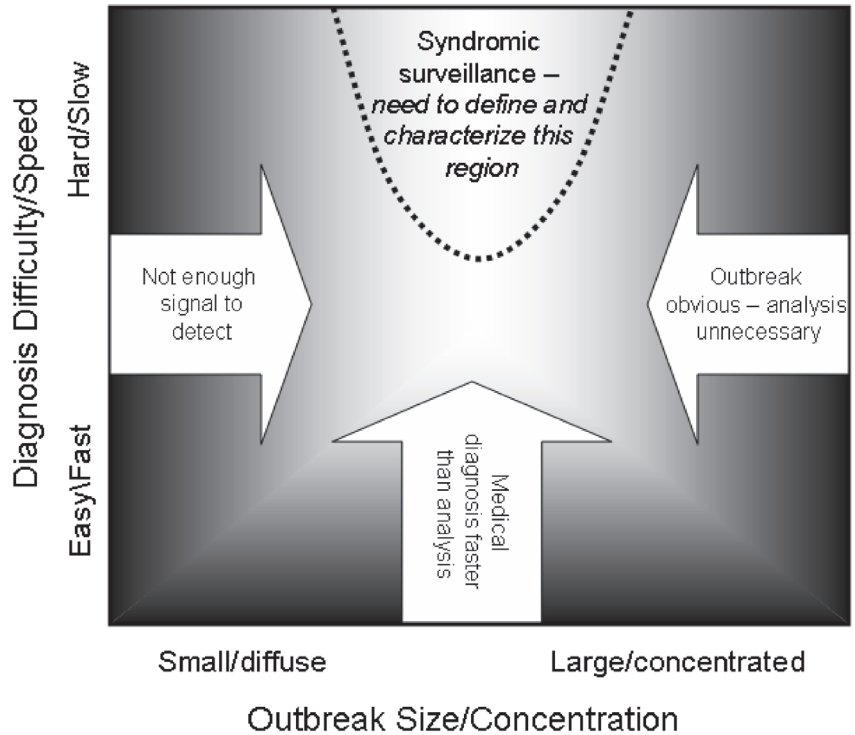


Figure 6. Is there a sufficiently large region in the diagnosis difficulty/speed vs. outbreak size/concentration region where syndromic surveillance systems (or other early event detection systems) would add value to justify current claims and hopes for such systems?

or skilled in other scientific and technical disciplines necessary for successful surveillance outcomes.

In addition, analytic biosurveillance findings are frequently the result of complex combinations of data inference and deductive reasoning that draw and fuse information from multiple sources. For example, the Department of Homeland Security has established the National Bio-Surveillance Integration System (NBIS), designed to analyze surveillance information from people, animals, plants, food, and the environment. This information will be monitored continuously and come from expanded and enhanced systems across the federal government and international sources. This creates formidable potential for situational awareness, but also exponentially expands the analytic complexity challenges (www.whitehouse.gov/omb/budget/fy2006/dhs.html). Thus, it also would be of great value to design systems and methods that allow users to think broadly about evidence in data and scientific logic.

What Is a Biosurveillance System Supposed To Detect, and How Fast?

A fundamental problem with electronic biosurveillance is that, even if we restrict its purpose to detecting outbreaks, no specific performance goals have been defined. Rather, there is just the overarching fuzzy purpose for such a system to signal disease outbreaks in a more timely and informative way than we could without it. Statisticians understand that numerous alternative hypotheses are more difficult to support with evidence from data than one narrow specific alternative. Real- or near real-time biosurveillance systems that are supposed to produce a signal for all possible types of disease outbreaks at any location or concentration at any time have fallen prey to “the mother of multiple alternative hypotheses,” and hence are doomed to suffer from decreased detection ability and high false signal rates.

Statisticians can help the public health community understand the util-

ity of being more specific about defining outbreak alternatives. That is, the better the public health community can define which diseases are of interest to monitor and how their outbreaks are likely to be manifested in the data being monitored, the better the analytical community will be able to define methods with the ability to best detect the specified outbreak with fewer false alarms.

Of course, it also is important to employ robust methods capable of detecting a variety of outbreak events, particularly with the prospect of deliberate bioterrorism attacks designed to be difficult to detect. But absent any clear statement of a set of specific objectives, the performance of detection algorithms in syndromic surveillance systems will be, at best, haphazard and, at worst, ineffective.

Developmental Challenges

How To Assemble and Interpret Electronic Biosurveillance Data?

Electronic biosurveillance system data are almost exclusively administrative or 'secondary use' data gathered for some purpose other than health surveillance. It is essentially available data that is logically thought to contain information that can be exploited for public health surveillance. It is frequently aggregated into or identified via arbitrary administrative boundaries, such as residence ZIP code or hospital visited. It is difficult at best, and often impossible, to link records for individuals across multiple hospital visits for an individual patient or between individuals' laboratory and pharmacy data. The data also can be subject to influences that have little or nothing to do with underlying health phenomenon, such as a sales promotion for an over-the-counter medication or "worried well" seeking medical attention in the aftermath of media coverage for some phenomenon. Often, measures are little more than aggregate counts over time by region.

Thus, there are many obvious research questions that need to be addressed: What do various counts mean? Can they be meaningfully converted into rates, and can and should those rates be risk adjusted for the underlying population? What is the best way to smooth across jurisdictional boundaries, particularly for small jurisdictional

areas? When should smoothing be used? For the majority of data that cannot be linked, should the "unlinked data source" problem be addressed, and, if so, how? The unlinked data source problem refers to the fact that, because most or all the syndromic surveillance data is unidentifiable, the extent of information overlap is unknown across data streams. For example, if a system uses over-the-counter sales, emergency room, and laboratory test data, it is not known to what extent the same people and their reactions to illness are manifested in the different sources.

Given that syndromic surveillance data are intended to be imprecise measures of some unobserved incidence of disease, what is their correlation with known outbreaks? That is, what is known about how much disease incidence signal is contained within the data? What types of nondiagnosis health-related data are most sensitive to the various types of disease outbreaks of interest? How should the data be assembled so as to maximize the likelihood of detecting an outbreak signal, or, conversely, what should be avoided or minimized that would tend to mask such a signal? How should data "lag time" and "time alignment" be addressed? Lag time is the time between an actual outbreak and the outbreak's manifestation in the data being monitored. Time alignment refers to the fact that an outbreak may manifest itself at different times in different data streams. For example, an outbreak may appear earlier in over-the-counter pharmaceutical sales data than for emergency department data.

What Is the Baseline State of Disease Incidence?

Industrial SPC methods are generally founded on the assumptions that much of the manufacturing process can be (at least broadly) controlled, particularly when the parameters of the SPC methods are first set to in-control conditions. Further, in an industrial setting, upon the signal of a procedure, the process can be evaluated to determine whether it needs to be adjusted or whether the signal was simply a false alarm.

The syndromic surveillance problem is not as simple. First, "in-control" is a nebulous concept as no control is actually exhibited over the process. For example, seasonal and other naturally

occurring and uncontrollable effects—including population changes—affect the observed outcomes. Second, it often is difficult and sometimes impossible to definitively identify various out-of-control conditions, even retrospectively. For example, identifying the start and end of the winter flu season in a particular region of the country is nontrivial and can be subject to debate. Isolating the underlying baseline disease incidence during this period is thus difficult at best. Further, from a bioterrorism detection perspective, the "normal" or baseline state includes naturally occurring "abnormal" events, such as the flu season.

This difficulty is additionally compounded in electronic biosurveillance systems with data that may be only loosely correlated with weak indicators of the actual underlying population health. It thus goes almost without saying that identifying an "out-of-control" outbreak state is made significantly more difficult when it is hard to characterize the "in-control" natural state. Thus, although SPC logic is applicable, an evolution of the methodology is in order to address the varying requirements implicit in the public health data and problem context.

What Are the Issues in Detection Algorithm Design?

In spite of the fact that what is known about how outbreaks will be manifested is limited, we can discuss some of the issues and challenges that need to be addressed in the design of effective detection algorithms. The first is to consider the temporal and spatial components of disease outbreaks. Effective algorithms will likely need to account for these two aspects simultaneously. Most of the industrial SPC methodology is only temporally oriented. In contrast, traditional clustering methods tend to ignore the temporal dimension. Kulldorf's SaTScan (www.satscan.org) is both spatial and temporal, but originally designed to be applied retrospectively on a static database (vs. repeated applications to a dynamic data stream in a surveillance context). This leaves a wide-open area ripe for new research.

Then, there is the question of the whether multiunivariate methods or true multivariate methods should be used, and under what conditions each is to be

preferred. As we discussed in the previous section, tradeoffs occur between the options, depending on whether an outbreak occurs across multiple jurisdictions or is best identified via changes in multiple data streams, for example. For multivariate methods, research is still required—at least for industrial SPC-based multivariate methods—to derive and characterize the performance of methods that are not direction invariant, meaning they can identify increases in disease incidence, but will not signal when decreases occur.

Most SPC procedures, including those previously described here, have been developed under the assumption that successive observations are independent. In industrial applications, this often can be reasonably well achieved by taking observations sufficiently far apart in time. For syndromic surveillance, data that exhibit characteristics such as seasonal cycles or other trends, such independence assumptions are more dubiously made. For example, many naturally occurring communicable diseases exhibit a seasonality component, such as a spike in cases in the winter. However, the timing, duration, and magnitude of these spikes often are not predictable.

The lack of a clear-cut baseline state combined with data that likely have a periodic component suggests methods that compare predicted rates from estimated baseline rates. Such methods are likely to require the use of a time gap in the data used for baseline estimation so data from an undetected outbreak does not contaminate the baseline estimate. However, it is not well understood how to design such time gaps into detection methods, what the tradeoffs are for longer vs. shorter gaps, and whether simpler methods without the gap can perform just as effectively.

How To Judge Algorithm Performance?

As it is difficult to characterize the normal or baseline state of disease incidence (and their associated biosurveillance system data measures) and the likely outbreak conditions have not been specified, it is rather difficult to evaluate and compare the performance of the various detection methods. Compounding this is a general desire by the public health community to see meth-

ods demonstrated on ‘real’ data. Yet, due to confidentiality and privacy concerns, there is a lack of general availability of such data to the research community. The result is a need for reasonably realistic pseudo-data that the public health community finds convincing but that, at the same time, is reproducible and can be released for general use. In order to generate such data, more fundamental research is required to better understand and characterize the various baseline states of disease incidence and the likely outbreak conditions.

Given such data, the research community must establish a common set of metrics to be used to evaluate the performance of the various detection algorithms. As we have discussed, the industrial SPC community uses average run length, though due to the likely autocorrelation in syndromic surveillance data and an inability to “reset” the process to an in-control condition (resulting in multiple signals over short periods of time related to one abnormal condition), the ARL measure may or may not be appropriate in this application. Other measures, such as average time between first signals, need to be rigorously defined and then accepted for widespread use.

Conclusions

Electronic biosurveillance systems are under development and being implemented around the country. These applications and the analytical methods used within them are motivated by a need for improved public health surveillance, not only for bioterrorism, but also to improve responsiveness to natural disease outbreaks (e.g., avian flu and SARS).

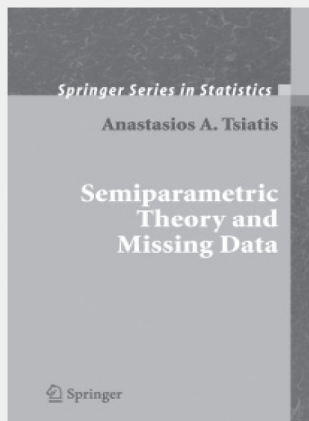
Initial system implementations have been driven largely by advances in communications and computer technology that now facilitate the real-time or near real-time collection and aggregation of health-related data. While in all these systems it is common to bring together various types of health and health-related data, exactly what these data are to be used for is still relatively undefined. Some efforts are focused on looking for changes in the rate of disease incidence, some are focused on changes in counts, some are focused on detecting clusters of disease events, and some are focused on detecting individual events. Each of these has different

implications for the types of outbreaks that will be best detected.

Because goals of biosurveillance are evolving in specificity, the current state of detection research is that each researcher defines his or her own assumptions and conditions and tests his or her algorithm under those conditions only. It is as if the choice of nail is dictated by the hammer at hand, rather than the best hammer being chosen to optimally drive the requisite nail. As the eighth report by the Committee on Government Reform, titled *Strengthening Disease Surveillance*, stated, “Successfully operating the elaborate, elegantly sensitive surveillance network of the future will require unprecedented levels of human skill, fiscal resources, medical information, and intergovernmental cooperation.” The statistical profession has much to offer and formidable responsibility in this endeavor.

Further Reading

- Fricker, Jr., R.D. (2006). “Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance.” *Advances in Disease Surveillance*, 1(3).
- Gerstman, B.B. (1998). *Epidemiology Kept Simple: an Introduction to Classic and Modern Epidemiology*. New York: John Wiley & Sons.
- Rolka, H.R. (2006). “Data Analysis Research Issues and Emerging Public Health Biosurveillance Directions.” In *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*. Wilson, A.; Wilson, G.; and Olwell, D.H. (Eds.). New York: Springer.
- Stoto, M.A.; Fricker, Jr., R.D., et al. (2006). “Evaluating Statistical Methods for Syndromic Surveillance.” In *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*. Wilson, A.; Wilson, G.; and Olwell, D.H. (Eds.). New York: Springer.
- Woodall, W.H. (2006). “The Use of Control Charts in Health Care and Public Health Surveillance.” *Journal of Quality Technology*, 38:1–16.



Semiparametric Theory and Missing Data

A. A. Tsiatis, North Carolina State University, Raleigh, NC, USA

Missing data arise in almost all scientific disciplines. In many cases, the treatment of missing data in an analysis is carried out in a casual and ad-hoc manner, leading, in many cases, to invalid inference and erroneous conclusions. In the past 20 years or so, there has been a serious attempt to understand the underlying issues and difficulties that come about from missing data and their impact on subsequent analysis. There has been a great deal written on the theory developed for analyzing missing data for finite-dimensional parametric models. This includes an extensive literature on likelihood-based methods and multiple imputation. More recently, there has been increasing interest in semiparametric models which, roughly speaking, are models that include both a parametric and nonparametric component. Such models are popular because estimators in such models are more robust than in traditional parametric models.

Introduction to semiparametric models.- Hilbert space for random vectors.- The geometry of influence functions.- Semiparametric models.- Other examples of semiparametric models.- Models and methods for missing data.- Missing and coarsening at random for semiparametric models.- The nuisance tangent space and its orthogonal complement.- Augmented inverse probability weighted complete case estimators.- Improving efficiency and double-robustness with coarsened data.- Locally-efficient estimators for coarsened data semiparametric models.- Approximate methods for gaining efficiency.- Double robust estimator of the average causal treatment effect.- Multiple imputation: a frequentist perspective.

2006. Approx. 400 p. (Springer Series in Statistics) Hardcover
ISBN 0-387-32448-8 ▶ \$84.95

Easy Ways to Order for the Americas ▶ Write: Springer Order Department, PO Box 2485, Secaucus, NJ 07096-2485, USA ▶ **Call: (toll free)** 1-800-SPRINGER ▶ **Fax:** +1(201)348-4505 ▶ **Email:** orders-ny@springer.com or **for outside the Americas ▶ Write:** Springer Distribution Center GmbH, Haberstrasse 7, 69126 Heidelberg, Germany ▶ **Call:** +49 (0) 6221-345-4301 ▶ **Fax:** +49 (0) 6221-345-4229
▶ **Email:** SDC-bookorder@springer.com ▶ Prices are subject to change without notice. All prices are net prices.