STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:513–529
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2032

Statistical issues and challenges associated with rapid detection of bio-terrorist attacks

Stephen E. Fienberg[1] and Galit Shmueli[2,*,†]

[1]*Department of Statistics, Center for Automated Learning and Discovery, Center for Computer and Communication Security, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.*
[2]*Department of Decision and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, U.S.A.*

SUMMARY

The traditional focus for detecting outbreaks of an epidemic or bio-terrorist attack has been on the collection and analysis of medical and public health data. Although such data are the most direct indicators of symptoms, they tend to be collected, delivered, and analysed days, weeks, and even months after the outbreak. By the time this information reaches decision makers it is often too late to treat the infected population or to react in some other way. In this paper, we explore different sources of data, traditional and non-traditional, that can be used for detecting a bio-terrorist attack in a timely manner. We set our discussion in the context of state-of-the-art syndromic surveillance systems and we focus on statistical issues and challenges associated with non-traditional data sources and the timely integration of multiple data sources for detection purposes. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: confidentiality; detection systems; ROC tradeoffs; privacy; record linkage

1. INTRODUCTION

Suppose that a group of terrorists exposed a capacity crowd of 15 000 in attendance at a college basketball game to an aerosolized form of anthrax through the arena ventilation system. How long would it take before those in attendance began to show symptoms? When would medical and public health officials begin to piece together the nature and extent of the exposure? Could an early warning help alert officials and thus save the lives of many of those exposed?

*Correspondence to: Galit Shmueli, Department of Decision and Information Technologies, 4361 Van Munching Hall, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, U.S.A.
†E-mail: gshmueli@rhsmith.umd.edu

Contract/grant sponsor: The Agency for Healthcare Research and Quality; contract/grant number: 290-00-0009
Contract/grant sponsor: National Science Foundation; contract/grant number: EIA-013884
Contract/grant sponsor: Army Research Office; contract/grant number: DAAD19-02-1-0389

 Accepted July 2004

The traditional focus for detecting outbreaks of an epidemic or bio-terrorist attack has been on the collection and analysis of medical and public health data. Although such data are the most direct indicators of symptoms, they tend to be collected, delivered, and analysed days, weeks, and even months after the outbreak. By the time this information reaches decision makers it is often too late to treat the infected population or to react in some other way. In this paper, we explore different sources of data, traditional and non-traditional, that can be used for detecting a bio-terrorist attack in a timely manner.

We begin, in Section 2, by describing the various aspects of the timeliness requirement, and its impact on the different stages of data collection, transfer, analysis, and decision making. At each stage there are different challenges that need to be addressed, and we detail these.

Several of the state-of-the-art syndromic surveillance systems under current development monitor real-time hospital or emergency room admissions. Although the data in such systems might be collected more frequently than in previous systems, the signal resulting from an outbreak tends to be too weak for detection. Solutions that have been suggested are: (1) aggregate the data temporally, (2) aggregate the data spatially, and (3) combine the collected data with those from other sources. The third solution, combining data from multiple sources, motivated us to explore the potential of monitoring non-traditional data and in particular grocery sales. In Section 3, we first discuss the usefulness of grocery purchase data by themselves for rapid detection of massive bio-terrorist attacks. The idea is to detect early signs of the epidemic before people arrive at medical facilities. We assume that self-treatment, which occurs earlier than medical treatment, manifests itself in grocery sales. For instance, following a massive anthrax attack we would expect an increase in sales of flu-like over-the-counter (OTC) medications as well as related groceries (e.g. orange juice). The next step is to integrate this system with the other medical and public health bio-surveillance systems.

In Section 4, we discuss the advantages and disadvantages of the different data sources and address the privacy and confidentiality challenge of combining data from these various sources.

2. SYSTEM AND DATA REQUIREMENTS FOR TIMELY DETECTION

Various medical and public health data sources have been routinely collected and monitored as part of efforts to detect outbreaks of an epidemic. These include visits to emergency medical services and sentinel practices, 911 calls and ambulance dispatch records, laboratory and mortality records, veterinary reports, and school or work absence records.

More recent efforts have been focused on developing a bio-surveillance system that will detect a large-scale localized outbreak resulting from a bio-terrorist attack involving known agents such as anthrax. The main improvement of such systems relative to existing systems is their ability to detect an outbreak much faster than public health officials have traditionally thought necessary. Wagner *et al.* [1] discuss the timeliness requirement in relation to the economic impact of an early detection system.

A rapid detection system requires special characteristics of the data that it monitors, as well as specific features from the monitoring system. Starting with the very first step of data

collection and data transferring, the initial requirements are:

1. *Frequent collection of data*. In many routinely collected medical and public health data the frequency of collection is as low as weekly (Microbiology lab reports in UK, Catchpole 2002), monthly, quarterly or even annually. No matter how fast the rest of the surveillance process works, such infrequent data will cause a delay that cannot be overcome in detecting outbreaks of a disease or a bioterrorist attack.

   Efforts to obtain frequently collected data have gone in two main directions: The first has involved efforts to facilitate and improve upon the collection of current traditional data. For example, the electronic surveillance system for the early notification of community-based epidemics (ESSENCE),[‡] which monitors patient data from military treatment facilities in a 50-mile radius of DC, records patient visits on a daily basis. The second direction has been to look for other sources of data that are collected frequently for purposes other than detecting outbreaks. Many companies have the resources and incentive to collect real-time or nearly real-time data. Reasons for such frequent data collection are marketing efforts in the presence of competition, billing requirements, inventory needs in a fast-moving market or for time-sensitive goods, and security reasons. Resources such as automated scanners and video cameras as well as other surveillance mechanisms allow for real-time data collection.

   Even if data are collected very frequently, they must be stored electronically to avoid delays in their transmission, and processed in ways to make them useful for detection purposes.
2. *Fast transfer of data*. Traditionally, data from medical and public health sources have been recorded and reported non-electronically in the form of oral, hand-written, or typed reports. This obviously creates not only a source of erroneous transmission, but also a serious delay in the flow of data. Non-electronic data requires an additional step of processing, which includes coding, editing, and storing [2]. Thus, electronic recording is essential to expediting the step of data transfer.

   Several projects that are aimed at rapid detection have tried to concentrate on improving the speed of data transfer by moving to electronic reporting. Bean and Martin [2] introduce a network for electronic surveillance reporting from public health sources.

   Another closely related issue is that of converting data from different information systems that differ between medical, public health, and other sources (hospitals, schools, pharmacies), or even within a given discipline (e.g. different laboratories) into a uniform format. The main challenges and costs involved in implementing an electronic reporting network are adopting new technologies, allocating funds for personnel (hiring, training, etc.), and allocating funds for the purchase and maintenance of hardware, software etc. In the implemented NY city surveillance system, one of the lessons learned was the need for pre-transmittal standardization and automated quality control.

   An example where data collection is relatively frequent, but the data transfer is slow is the ESSENCE system (see above) which records patient visits on a daily basis, but has a 1–3 day delay in data transfer [3].

[‡]http://www.geis.ha.osd.mil/GEIS/SurveillanceActivities/ESSENCE/ESSENCE.asp

An example of an electronic reporting network where both data collection and transfer are rapid is the real-time collection of ER visits used by the real-time outbreak and disease surveillance (RODS) system,[§] which has been deployed in Western Pennsylvania for 2 years. A key feature of RODS is that it receives data directly and without delay from computers in emergency departments and hospitals. The RODS laboratory also receives data from other sources such as retailers. These data are received with a maximum delay of 24 h. The NYC syndromic surveillance system is another example.

The National Electronic Disease Surveillance System is a CDC initiative to 'facilitate the electronic transfer of appropriate information from clinical information systems in the health care industry to public health departments, [to] reduce provider burden in the provision of information, [and to] enhance both the timeliness and quality of information provided' (from `http://www.cdc.gov/nedss/`).

In addition to these two requirements for timely collection and transmittal of data, three major features of data that are essential for timely detection are:

1. *Early signature of the outbreak*. The main idea behind syndromic surveillance and the collection of non-traditional data is that an outbreak will manifest itself in such data before it does so in traditional medical and public health data. This assumption which is pivotal to syndromic surveillance, is based on the fact that people tend to pursue self-treatment before seeking medical assistance. Syndromic surveillance usually uses data that are routinely collected for other purposes [4]. The idea behind syndromic surveillance is that it might detect the signature of a disease a day or even a week before the disease itself becomes apparent. Examples of data that carry an early signature of an outbreak are sales of pharmacy and over-the-counter medications, searches on health websites such as WebMD, and bio-sensor measurements that are continuously taken from individuals in the relevant area.
2. *Sufficient amounts of data*. In order to detect an outbreak rapidly, we must have enough information. If the number of observations at each time period is too small it will be hard to distinguish between ordinary variability and an outbreak. Several researchers have shown that a lack of sufficient data might lead to under-detection of real outbreaks. One artificial way to overcome this problem is to aggregate data over several time periods [5] or aggregate over a larger geographical area. Aggregating data temporally will automatically slow down the speed of detection, however, since we must wait several time periods until there is sufficient data for the surveillance system to examine. In spatial aggregation the outbreak signal is dampened, and a large-scale outbreak in a geographically limited area would not be noticeable among the 'no-outbreak' data that is added from other regions.

   A major issue for data transmission is whether or not the data are aggregated or are available at the individual level. Aggregated data are simpler to transmit and raise limited privacy issues. Individual level data are more voluminous, raise numerous privacy concerns, and may not be easily merged because of differences in forms of identifiers

[§]http://www.health.pitt.edu/rods/

in diverse systems, as well as errors in spelling, data entry, etc. For example, we know from experience that the rate at which names such as *Shmueli* and *Fienberg* are entered into official systems incorrectly is high! We discuss these issues of privacy and record linkage in more detail in Section 4.

3. *Local, not regional or national data.* The type of outbreak that we are trying to detect is a large-scale localized outbreak, resulting from a bio-terrorist attack in a limited geographical area. Thus, the most informative data are those collected in the infected area. Aggregating data spatially can significantly reduce the chance of detection, since data from unaffected areas will mask the outbreak signature and it will take longer to notice the slight change. Collecting localized data is therefore not only beneficial for improving sensitivity, but also for timeliness of detection.

In comparing these elements in different data sources, it is clear that the signature of an outbreak will be earlier in non-traditional sources such as pharmacy and grocery sales than in traditional medical and public health sources. An even earlier signature would potentially exist in embryonic systems that are not yet available. A few examples of such sources are Web browsing at medical websites such as WebMD, automatic body tracking devices (BodyMedia), and Biosensor data [6]. However, it is still unclear how these data can be utilized in the form that they are currently collected.

Even if the above requirements from the data are met, the detection system might not supply timely decisions if there is a bottleneck in the analysis or output stage. These are what we refer to as 'system requirements':

1. *Immediate analysis of incoming data.* In order to be able to analyse the data quickly for outbreak signals, we must be able to store incoming data in a format that the detection algorithms can use. This means that resources are needed for quick storage. Next, in order for the detection algorithms to run quickly, they must be based on efficient computation. These two features of fast storage and analysis are especially critical when dealing with massive data sets.
2. *Immediate output*. After the statistical analysis is completed the detection system should output an operational decision-making conclusion. Since the users of the system will usually not be statisticians, the output must be in a user-friendly format which can be easily understood and transferable (graphs, reports). Of course this output should be immediate and not delay the process of decision making.

The requirement of timeliness is in addition to the two usual requirements of a high true-detection rate and a low false-alarm rate. Current systems tend to have a high true-detection rate, but at the cost of a high false-alarm rate. One of the most developed syndromic surveillance systems is the one used by the New York City Health Department. They collect and monitor large amounts of traditional and non-traditional data (ER admissions, 911 calls, Pharmacy sales, etc.) that are reported on a daily basis [4]. However, their system reports a false alarm almost every week (*New York Times*, April 4, 2003). For each alarm, experts examine the situation and assess the chance that it is a true signal. All systems can be calibrated to have less false alarms at the expense of a slower true detection rate, and vice-versa. Although the risk of not detecting a true outbreak should be minimized, the cost and handling of false alarms must be taken into account.

3. OUTBREAK DETECTION USING GROCERY SALES DATA

We focus here on the investigation of and system development for a single source, grocery sales, to emphasize the complexity of the detection problem. Although the data seem promising for fulfiling many of the timeliness requirements, they pose structural and statistical issues that must be addressed.

Our claim is that grocery sales data are potentially valuable for rapid detection. They meet many of the requirements that we listed in Section 2. First, they are recorded electronically in real-time, which is the time of purchase, using scanners. Purchase data are often recorded and stored at very short time intervals, as short as minutes. This means that the information is practically real-time and can thus allow for rapid detection. Technically, they can then be transferred at an aggregate level of choice. It might seem that the optimal level for timely detection would be the highest level of detail. However, there are several issues to consider that can offset the timeliness advantage: First, high-detail data take longer to transfer and manipulate for both technical reasons as well as for privacy and confidentiality concerns that they raise. When data are at a high level of detail (e.g. at the basket level, where data include entire records for each basket of products that was purchased by a consumer at a certain time) and include identifiers (for the purpose of combining data sources), the data supplier might refuse to disclose the identifiers, or even the unidentified data at a high level of detail. See the related discussion in Section 4. Another consideration is the amount of noise in the data. Very frequent data (e.g. the sales every 10 min) will include so much variation due to extraneous sources that the noise might mask a real outbreak. In our project the data were available immediately at an aggregated level where we had the daily sales for each item, and only later on did we obtain basket-level data on an hourly basis. The two types of formats require different statistical methods since they are not only on different time scales but are also item-aggregated vs basket-level.

Second, we believe that sales of OTC medications and relevant groceries and their combinations represent an early signature of an outbreak. Although these data are usually collected for other purposes such as marketing, pharmacy data have been used previously for public health related investigations. Subramanyan *et al.* [7], for example, used automated pharmacy dispensing data to examine treatment and management of 45 cases of active tuberculosis in New England between 1992 and 1996. They mention that such records can also be used to identify cases of tuberculosis unknown to the public health system.

Third, purchase data are often localized and are therefore useful for detecting large-scale outbreaks in a relatively small area. Furthermore, when the data are collected from a chain of stores, as in our case, purchase location can be narrowed down to smaller geographical regions, based on the specific store location, even if purchasers' addresses are unavailable for privacy reasons.

Finally, scanner data are very rich in detail and can be as detailed as basket level. The amount of data is usually huge, which means that statistical and data mining methods can be more powerful than when analysing sparse medical or public health series. On the other hand, our experience with working with hourly data was that the processing time for extracting data in a useful format, and carrying out even the simplest calculations was prohibitively long. Thus, there is a potential value in using highly frequent, detailed sales data but the challenge of developing efficient methods for handling such data is real.

3.1. Designing a statistical framework

In order to use very frequent, highly detailed, symptom-related sales of OTC medications and grocery items within a detection system requires different statistical tools and raises different statistical issues relative to medical or public health data. The fine time scale means that the dependence between sales within neighbouring periods of time cannot be ignored. Thus, popular tools such as simple control charts are inappropriate. Secondly, in sales data the ratio between signal and noise, when the purpose is detecting a large-scale bio-terrorist attack, is much smaller than in public health series. Sales data include many factors other than the epidemic-related symptoms. Thus, we require a large amount of data and very powerful statistical methods for differentiating between relevant and irrelevant changes in sales. Next, we discuss a framework of how to design a statistical detection system, and we point out the challenges at each step.

1. *Decide which items to monitor*. In determining which grocery items and OTC medications to track, for detecting an outbreak following the release of a specific bio-agent, epidemiological expertise is necessary, but not sufficient. Each of the series pointed out by the medical experts must be examined from a statistical point of view, to see whether it contains useful information. Some series, for instance, are sales of high-volume items that would not change drastically enough even in the event of an outbreak. It is important to note that the items to monitor can include not only items that are expected to increase/decrease following an outbreak, but also items that would remain at their normal level, but which would help distinguish the outbreak of interest from a symptom-wise similar one. For example, anthrax and flu lead to similar symptoms, except that anthrax does not lead to a runny nose. Thus, monitoring the sales of Kleenex or nasal medication could help distinguish between the two.
2. *Model the* '*no-outbreak*' *baseline of sales*. The next step is to learn the sales behaviour of the items of interest when there is no outbreak. This requires a sufficiently large amount of preliminary data that are **known** to be void of the outbreak of interest. This is, of course, outbreak-specific, so if we want to detect, for instance, anthrax we require an 'anthrax-free' period of sales, while a detection of an influenza outbreak would require an 'influenza-free' period. Since 'anthrax-free' periods will most likely include influenza seasons, we want the baseline model to 'learn' what and when influenza looks like in the sales, and incorporate it as a 'non-outbreak' situation. At this point the knowledge and feedback from epidemiologists and public health experts is necessary in order to verify that the period is indeed free of the relevant outbreak.

   In order to minimize the 'noise', or irrelevant fluctuations in sales over the baseline period, we must account for total sales, marketing efforts, promotions, etc. In order to account for such factors, the data must include information such as regular price, reduced price (if there is a promotion), overall sales, etc. These data might not be available, however, and even when they are it is unclear how and to what extent they affect sales of items relevant to symptoms. For example, cough medication sales are probably less sensitive to promotions relative to orange juice or soup. Furthermore, the combined effect of promotions and seasons (e.g. the promotion of an antihistamine brand during allergy season) can create further variability in sales. Thus, whether the complete information on

sales is given or not, the challenge of 'cleaning' the data from factors that are irrelevant to an outbreak is a major challenge.

3. *Simulate an outbreak signature in sales data*. Several authors have researched the signature of anthrax in traditional data sources. Brookmeyer *et al*. [8] studied the Sverdlovsk data in order to learn about the progression of anthrax and the mortality rate.

   Wein *et al*. [6] developed a mathematical model that includes disease progression after an airborne anthrax attack. In their intervention or emergency response model they assume that intervention begins 48 h after the attack. Their model relies on information from local hospitals, neighbourhood emergency health centres, etc. In related efforts, Wagner et al. [1] discussed the footprint of an anthrax outbreak in medical data, and Pavlin [9] described the difference between the epidemic curve of a disease versus a bioterrorism attack.

   What would the signature of a large-scale airborne anthrax attack look like in pharmacy or grocery sales data? One of the issues that involves the most uncertainty in monitoring non-traditional data is the manifestation of the outbreak in the data. Since the data are collected for other reasons and do not directly measure the outbreak results, and since there are no sales data available that include a period when a mass bio-terrorist attack such as anthrax occurred, the usual ways of learning and constructing the outbreak signature do not always work.

   Our approach in such cases is to build a typical 'anthrax signature' in sales data by collaborating with epidemiologists, public health officials, and marketing experts. A group of such experts can come up with the most likely signature. Using this method for the case of anthrax, we constructed a signature that has a linear increase in sales of medications that are aimed at symptom-relief such as cough medicine, but no increase in nasal decongestants (since the absence of this symptom in anthrax differentiates it from influenza).

4. *Test system for real and false alarms*. From the extensive literature on receiver operating characteristic (ROC) curves [10, 11] it is known that when the base rate of the quantity of interest is very low, there will be a high proportion of alarms, both true and false ones. In our case, not only is the (hypothesized) base rate low, but we do not even know the exact form of the outbreak signature in grocery data. Thus, we cannot determine a threshold for alarms that would yield required alarm rates, or vice versa: we cannot compute the exact alarm rates for a given threshold. The gap between the proposed simulated signature and the real one, coupled with the low base rate, will inevitably lead to high alarm rates, which are not directly measurable before the outbreak occurs.

   Our suggested statistical method for selecting a threshold for alarm and estimating the true and false error rates is by simulation: we insert the simulated outbreak signature into real sales data at various time periods within the data range, thereby simulating the outbreak of interest situation multiple times. The number of false and real alarms can then be estimated from these simulations, and used as gross evaluations of the real rates.

5. *Develop a roll-forward algorithm*. Since the data arrive point by point (either a single measurement such as overall daily sales of cough medication, or a multivariate measurement such as daily sales of a multitude of items), we would like an algorithm that is able to integrate all the previous data and use it to decide whether the new data point indicates an outbreak or not. This feature is called 'role-forward'. There are several issues

involved in developing a role-forward tool:

- We need data to initialize or 'start-up' the system. Some algorithms require more data than others. However, this is usually not a problem, since some historical data are usually available.
- Searching for a future outbreak is different from trying to identify an outbreak in historical data (e.g. identifying TB cases in historic pharmacy data, [7]. This issue of 'online' vs 'offline' data and analysis is similar to the one encountered in quality control, where there are different ways of analysing control charts, depending on their detection purpose. Instead of running through a given series once, we assess every new data point. This means that the computation is repeated after each data point is added.
- Since the computations are redone at every entry of a new data point, it is essential that the algorithm be efficient, i.e. will be fast and computationally cheap. This requirement gives priority to some statistical methods over others.

Another requirement is that the detection system be fully or almost fully automated. This means that the statistical tools used should not be too customized to a specific data set, to the extent where it needs too much manual tweaking and adjusting. It should also be flexible enough for handling different outbreak types. Thus, once the signature of interest is constructed and the data set selected, we should be able to 'feed' them to the system with minimal adjustment. From our experience it is not possible to reach a fully automated system, but a careful selection of methods and algorithms can minimize the need for intervention. The cost, obviously, is an overall reduction in discrimination power.

The issue of 'user-friendliness' of the system is important. For non-statistician users, who need to reach operational decisions based on the system output considerations of software availability and informative output are major issues. Even if there are statistically literate experts on the decision-making team, the meaningfulness of the output and the ability to make changes to the system will depend on how 'transparent' the algorithm is, and whether it is based on tractable or black-box methods.

Finally, the ideal system would have the ability to be integrated with other systems, and to integrate several data sources.

3.2. Implementing the framework for grocery sales

To illustrate some of the statistical issues and challenges that we described in the previous sections, we use the example of the detection system that we developed, which uses grocery and OTC medication sales at a major retailer with many branches in the Allegheny County, PA area, between August 8, 1999 and January 31, 2001. Goldenberg *et al.* [12] gives overall description of our method and approach and Goldenberg *et al.* [13] provides technical details.

Grocery and over the counter medication sales data are typically very large and rich, including information on each purchased item and in many cases include customer information which is obtained through affinity cards. Our data were available in two formats: We began with daily-aggregates of the sales of several groups of medications (e.g. cough medication, nasal decongestants, cold medication). A few months later we obtained more detailed data—basket-level data on an hourly basis. Thus, the first data set was aggregated temporally and

over baskets, whereas the second data set essentially had complete detail. In both cases the data sets were large and covered a period of more than a year; they were local to the Pittsburgh vicinity; and were stripped of identifiers for privacy reasons, both at the individual customer level and in terms of the location of the store where the purchases were made. The absence of these identifiers restricts the usefulness of the data.

To explore the potential of using these data to detect a massive large-scale airborne anthrax attack, we first looked at detecting influenza, which leads to similar symptoms as those that follow exposure to anthrax. Several flu outbreaks occurred during the period of our data and sales of some relevant medications such as cough medication appeared to peak a few days earlier than the peaks in the traditional public health and medical data. Since most of the traditional data were available only on a weekly basis, however, the comparison was not clear cut. Our next choice was the time scale of the data. We decided to work with daily data for several reasons: First, the initial data that were available to us were daily aggregates. Second, experts on the progression of anthrax talk about a daily progression. Third, other systems were aiming at obtaining daily data (which would be an improvement over the weekly scale). Fourth, from a statistical point of view, it seemed that even at this level there was enough information for detecting abnormalities whereas the hourly data, which we obtained later on, were extremely noisy. Finally, by analysing daily data one automatically adjusts for daily trends (e.g. more drugs sold during the day than during the night).

Anthrax-related symptoms are very similar to flu-related symptoms with one major difference: flu creates nasal congestions, whereas anthrax does not. Thus, to distinguish between the two using sales of medications, we would need to monitor items that treat a symptom that is *unrelated* to anthrax, i.e. nasal decongestants, Kleenex, etc. Next, we used statistical expertise to assess whether the sales series of the relevant items were informative. Figure 1 describes the daily sales of three medication subgroups. It can be seen that nasal decongestion medications are sold at a much lower volume than cough and sore throat medications, and that they fluctuate much less. This might be related to the fact that nasal decongestant medication is used for treating allergies, which occur year round. Thus, it seems that it would be harder to detect a change or a non-change in nasal decongestant medications. This led us to focus on the sales of cough medication with the initial data. On the other hand, for the basket-level data set we chose to monitor purchases of various symptom-relieving medications and items, so in addition to the three medication subgroups we monitored several other medications but also grocery items such as tissues, orange juice, and soups. An increase in a combined purchase of these could then be indicative of a flu or an anthrax outbreak (depending on the presence of tissues and nasal decongestants in baskets with other anthrax symptom-relief items).

After deciding on the data to be used by the monitoring system, we aimed to establish the 'no-anthrax' baseline. Fortunately, during this 1999–2001 period there was no known anthrax outbreak in this area. Nonetheless, as Figure 1 indicates, the sales of cough medication have widely varied patterns: A seasonal effect, with winter sales higher and more chaotic than summer sales, a weekly effect showing higher sales during weekends, peak sales on holidays, and low sales on days when many stores are closed (e.g. Easter). There are two main factors that influence the sales of cough medication other than an anthrax attack: General patterns of sales at grocery stores, and outbreaks of diseases such as influenza, where cough is a major symptom as well.

With the data that were available to us we were able to partially account for fluctuations in overall sales but not promotions. For cough medication, however, we felt that sales would

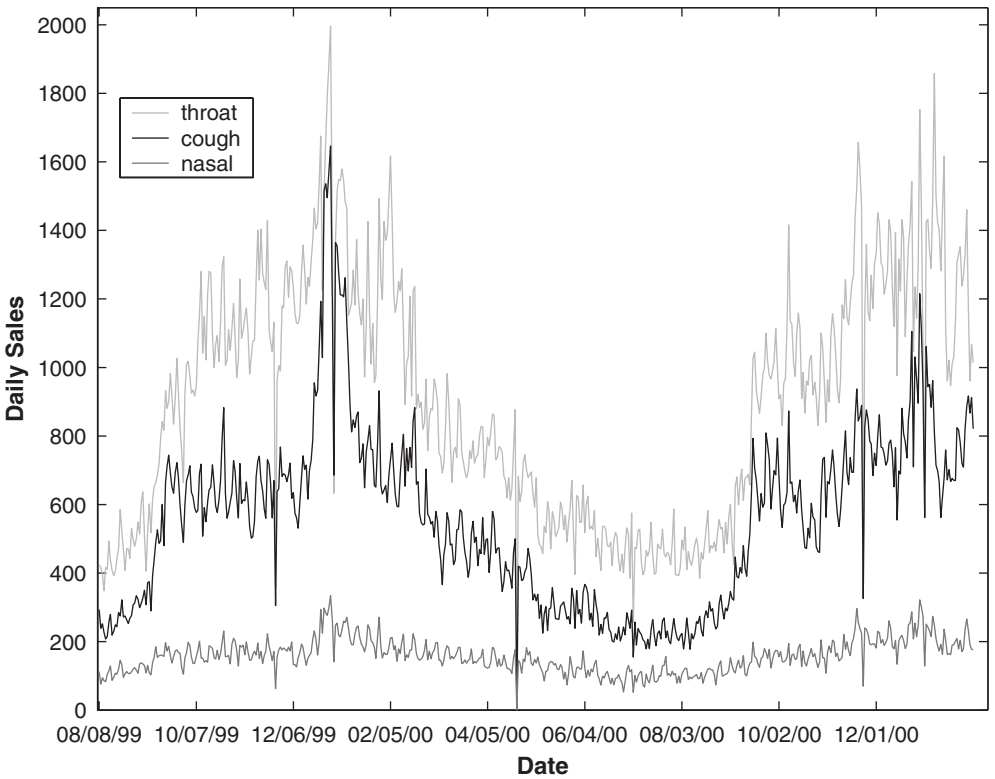

Figure 1. Sales for three OTC medication subgroups from 8/1999–1/2001.

be relatively insensitive to promotions. Since there was still much variability left in cough medication sales, we used smoothing methods to reduce the natural variability, then estimated this variability. All the statistical methods that we applied had trouble 'learning' the extreme changes in sales on holidays. Thus, the algorithm always signaled a false alarm on these days. However, since holidays dates are always known in advance they could be entered into the system ahead of time, and the system could then attach a comment to an alarm that is raised on a holiday.

We sought assistance from epidemiologists to evaluate how an anthrax attack would manifest itself in cough medication sales. The simulated footprint spanned a 3-day period, where sales were expected to increase linearly. Taking into account the quick progression of anthrax, only the first three days after the attack are simulated since detection on later days is too late and thus is practically useless.

In selecting the statistical methods and tools that were implemented in the detection system, we preferred ones that require less tuning, and that are more easily automated. For example, the series of cough medication sales is not a stationary series, and therefore requires several manipulations before an ordinary autoregressive moving average (ARMA) model can be fitted. Also, an ARMA model would require the user to determine the order of the

autoregressive and moving-average components. Moreover, if we were to use a different series altogether, we would have no guarantee that it could be transformed to a stationary series, and thus an ARMA model would not be appropriate at all. Instead, we used wavelets, which can be applied directly to non-stationary series, and do not require much specification from the user. Wavelets were chosen for two additional reasons: The algorithm that carries out wavelet analysis is very efficient and does not require much computing time or power, so it is suitable for a role-forward algorithm. In addition it is a tractable method (unlike neural networks, for example), which make it more appealing for analysis purposes. Wavelet analysis, however, requires an initial data set, which should not be a problem in such a setting.

The next challenge was to set the threshold for alarms: since we did not have any data that included an actual anthrax attack, we used the data that we had but with the simulated anthrax signature added to it. To avoid over-fitting by setting a threshold according to the number of real and false alarms for these data, we looked at the model residuals and used a control chart type argument about approximate normality, and used it to differentiate between natural variability in sales and anthrax-related variability. This allowed us to apply a 3-sigma type threshold. Another issue was in determining the definition of a detection. Taking into account the 3-day progression of anthrax, and the 3-day anthrax signature we defined a successful detection as exceeding the threshold at least once within 3 days of the simulated attack. Borrowing ideas from statistical quality control was appealing for another reason: many other detection systems that are used by the Centers for Disease Control and Prevention are based on using control charts for monitoring medical and public health data. Thus, the users of our system would be familiar with the concepts of a 3-sigma threshold, natural variability, etc. and would be able to interpret the system's output (in the form of a graph of the series with a threshold) without errors or delay.

For the basket level data, the choice of statistical methods and data mining tools was even more critical due to the massive number of baskets. Even when we restricted ourselves to baskets with one or more of 50 pre-selected products, we had 200 000–500 000 baskets a week. To determine the 'no-outbreak' baseline we needed to evaluate the relation between the 50 products. How likely are we to find different combinations in purchasers' baskets? We applied the method of association rules to these data in order to find the most likely combinations (or 'rules'). The resulting number of combinations was very high and required us to use two additional steps of thresholding in order to find the 'most unexpected' combinations of products relative to what we would expect if there were no dependence between the products of that combination. We ended up with pairs and triplets of products, which we then tracked to see which repeated over time.

The resulting most common triplets were surprising: when looking at the most likely combination of 3 flu-related products out of the 50, 'soup triplets' (including condensed soup, ready-to-eat soup, etc.) were the most stable combination that repeated week after week. This anomaly resulted from several marketing related factors: First, soup is purchased at a much higher frequency than health-related items. Second, the variety seeking in soups is much higher than, say, in laxatives. So people are more likely to purchase several different soup items in a single purchase. Finally, the classification of grocery products and OTC medications have different levels of detail which makes their comparison complicated. Thus, when trying to merge two types of data even within the same data set and from the same data source, a serious challenge arises.

4. COMBINING DATA SOURCES: BENEFITS AND CHALLENGES

Various ongoing surveillance projects now collect data from several traditional and non-traditional sources and attempt to use these data for detection purposes. Two examples are the RODS laboratory, which collects data on ER and hospital visits in Pennsylvania and Utah, and OTC healthcare products in Pennsylvania; and the ESSENCE II system which includes data from military and civilian outpatient visits, OTC pharmacy sales, school absenteeism and animal health data.

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. For example, in various data systems that collect grocery purchases and pharmacy purchases, over 80 per cent of grocery data can be linked to families through affinity cards, whereas pharmacy purchases may also be labelled using the same or similar affinity cards as well as by actual names and possibly other identifiers on individual prescriptions. Linking such sources is a serious statistical issue, unlike the linkage of hospital admissions and ER records, which are directly linked to the person admitted. Other non-traditional non-public-health data include school absence records that contain names and related identifiers and records of 911 calls in New York City, which include the patient's age, gender and ZIP code (together with their chief complaint). The identifiers we have listed above may not even, in principle, be the same, and matching names and fields, especially in the presence of substantial recording error, poses substantial statistical challenges. We discuss some of these challenges here.

4.1. Record linkage methods for producing integrated data files

There is a growing literature on record linkage emanating from a seminal paper by Fellegi and Sunter [14] and a more recent paper in computer science by Monge and Elkin [15]. Jaro [16] gives details in a public health context, and Bilenko *et al.* [17] provide more recent references on the subject. One can think of virtually all current methods as name-matching approaches as involving either a set of 'match features' or 'string distances' to be used to pair entities in two different data bases. To merge data from more than two lists, what is typically done is to take the lists in pairs and then use ad hoc methods to resolve inconsistencies [18]. What we need are formal extensions of existing approaches for the combination of multiple lists which also allow for missing identifiers in some lists.

4.2. How combined sources can be used for monitoring

One could attempt to avoid the record linkage problem by independently and simultaneously monitoring the separate sources, either using individual or aggregate data. This is the simplest strategy to implement and raises the fewest concerns regarding privacy and confidentiality. However, simultaneous monitoring has the danger of an inflated false alarm rate due to multiple testing, and thus methods that use this approach must compensate for it [19].

Another option is to track different series intensively but sequentially, where signals from early systems trigger further data collection or more likely intensive analyses of other series. Researchers at the University of Utah have developed such a system, where a signal from an influenza electronic reporting system elicits an active collection of lab cultures. An alternative 'hierarchical signalling' approach is one where signals from early systems alert later ones. In our project to extract information from grocery and OTC medication sales regarding a possible

anthrax attack we followed this approach, and thus presumed that a signal from the grocery and OTC medication sales would trigger alertness of public health and medical surveillance systems.

As statisticians we clearly favour a third approach that would attempt to produce merged records for individuals and/or families that could then be used with multivariate prediction models and would then hopefully have greater sensitivity to early signs of a bio-terrorist attack. The challenges here result not only from the complexity of such multivariate modelling, but also from the often substantial measurement error associated with less than perfect record linkage, and the privacy and confidentiality concerns raised by the merger of data from different record systems. Coping with privacy and confidentiality concerns associated with individual data systems is difficult enough, but concerns associated with merged record systems are typically far more problematic.

4.3. Privacy and confidentiality issues

The kinds of 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 'deidentified'. Similar legal restrictions apply to prescription information from pharmacies. Other public and semi-public 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 are typically the property of the commercial interests that are weary of sharing data in individually identifiable form even if there are no legal strictures against such access.

In light of such a situation how might we proceed to consider the possibility of a fully integrated early warning data system that was capable of merging records of such diverse forms across systems at the family or even individual level? First, the new HIPAA privacy and confidentiality standards are intended to balance privacy protections with the public responsibility to support such national priorities as protecting public health. Thus partnerships with diverse medical and public health organizations are not necessarily precluded by law, and in many cases may be specifically exempted.

Record linkage methods require useful and accurately recorded identifiers to produce accurate merged records, but once the merged files have been created there is not necessarily an analytic need for identifiers when the data are used for surveillance purposes. On the other hand, variables included in individual data sources may not pose a disclosure risk but when they are combined with other variables as part of an integrated data base they may pose a substantial risk. Doyle *et al.* [20] and Fienberg [21] introduce the methodological issues associated with confidentiality and disclosure limitation.

We see the following major challenges and issues in this domain for statistic research and methodology:

- How can we gain access to data from diverse sources subject to varying confidentiality rules and forms of privacy protections, especially when the data come from non-traditional sources?

- Can we separate out a neutral function for 'statistical' record linkage from the analytical detection function of surveillance systems? Such a trusted third-party may be essential to the production of 'real-time' updated files because identifiers are essential to rapid updating.
- Can we create merged files useful for surveillance purpose by different entities that will not allow for re-identification and linkage by the 'owners' of the separate originating data sources?
- Within a confidentiality framework that allows for the tradeoff between disclosure risk and data utility [22], should we be willing to tolerate greater risks of disclosure with demonstrable gains in utility of data for surveillance purposes?

These and other confidentiality issues require considerable technical attention as well as careful arrangements with those responsible for the data. Finally we offer a cautionary note: Experience suggests that statistical measures that people believe are sufficient to protect data from attack by an intruder may not in fact be so. Careful empirical testing of any such effort is a crucial component of work to protect confidentiality.

5. CONCLUDING REMARKS

The United States has increasingly focused on the role of surveillance systems in the detection of disease outbreaks as well as for the early detection of bio-terrorist attacks. While the development of such systems had been underway for several years, following the events of September 11, 2001 and the ensuing localized anthrax attacks [23], efforts and attention have been intensified. Although there have been a number of small-scale bio-terrorist attacks, we have yet to see a large-scale bio-terrorist attack, e.g. one involving anthrax. The two types of attack are different in nature, and our focus in this paper has been on large-scale attacks only. The main goal of timely detection of a large-scale bio-terrorist attack is to avoid a situation where a large proportion of the population is stricken before the medical and public health systems can respond.

The notion of being able to link multiple data sources from public health and medical sources on the one hand, and non-traditional sources such as medication and grocery sales on the other hand is key to this paper. In addition to the usual criteria of high true-alarm rate and low false alarm rate, the *timeliness* of any surveillance system using such data is a central concern. This poses requirements on the data collection and data transfer phases, as well as on the detection system, which should be able to handle large amounts of data and to carry out sophisticated analyses efficiently in close to real time.

In this paper we focus on the many statistical issues that surround the development of such detection systems, only some of which have received proper attention to date. While our own experience with such systems has been limited [12, 13], we have attempted here to address a broader set of relevant statistical and policy issues. In Section 2 we described the various aspects of the timeliness requirement, and its impact on the entire process of detection, from the collection of data through the analysis and until the decision making. We focused on the challenges that arise in this context and discussed the progress on related systems to date. Section 3 illustrates our efforts to explore the use of grocery and OTC medication sales for

early detection purposes, taking into account the different constraints and requirements that are associated with timely detection.

One critical challenge in the development of surveillance systems that involve diverse data sources is linkage of individual or family-level data from record systems that were designed for different purposes and which use different forms of identifiers. Privacy and confidentiality are concerns not only for systems based on single data source data, but also for the merged system, and control of and access to this merged system may be a highly contentious issue. In Section 4 we addressed some of the issues and challenges arising in this domain.

ACKNOWLEDGEMENTS

The discussion in Section 3 is based on collaborative work with Anna Goldenberg who worked on the daily aggregated OTC medication sales data, and Dunja Mladenic, who worked on the basket-level data. The authors would like to thank Martin Kulldorff for his useful comments and suggestions.

REFERENCES

1. Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, Caruana RA, McGinnis LF, Deerfield DW, Druzdzel MJ, Fridsma DB. The emerging science of very early detection of disease outbreaks. *Journal of Public Health Management and Practice* 2001; **7**:51–59.
2. Bean NH, Martin SM. Implementing a network for electronic surveillance reporting from public health reference laboratories: an international perspective. *Emerging Infectious Diseases: Perspectives* 2001; **7**:773–779.
3. Lee J-E. Analysis of a health indicator surveillance system: its ability to detect annual influenza activity for the 1999–2000 and 2000–2001 seasons compared to traditional surveillance systems. *International Conference on Emerging Diseases*, Atlanta, GA, 2002, http://www.cdc.gov/iceid/webcast/surveillance information.htm
4. Mostashari F. BT surveillance in NYC. *International Conference on Emerging Diseases*, Atlanta, GA, 2002, http://www.cdc.gov/iceid/webcast/surveillance.htm
5. Reis BY, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. *Proceeding of the National Academy of Sciences* 2003; **100**:1961–1965.
6. Wein LM, Craft DL, Kaplan EH. Emergency response to an anthrax attack. *Proceedings of the National Academy of Sciences* 2003; **100**:4346–4351.
7. Subramanyan GS, Yokoe DS, Sharnprapai S, Nardell E, McCray E, Platt R. Using automated pharmacy records to assess the management of tuberculosis. *Emerging Infectious Diseases: Research* 1999; **5**:788–791.
8. Brookmeyer R, Johenson E, Bollinger R. Modeling the optimum duration of antibiotic prophylaxis in an anthrax outbreak. 2003, unpublished manuscript.
9. Pavlin JA. Epidemiology of bioterrorism. *Emerging Infectious Diseases* 1999; **5**:528–565.
10. Green DM, Swets JA. *Signal Detection Theory and Psychophysics*. Wiley: New York, 1966 (reprinted Peninsula Publishing Co.: Los Altos, CA, 1988).
11. Swets JA. *Signal Detection Theory and ROC Analysis in Psychology and Diagnosis*. Erlebaum: London, 1996.
12. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceeding of the National Academy of Sciences* 2002; **99**:5237–5240.
13. Goldenberg A, Shmueli G, Caruana RA. Using grocery sales data for the detection of bio-terrorist attack. 2003, unpublished manuscript.
14. Fellegi IP, Sunter AB. A theory for record linkage. *Journal of the American Statistical Association* 1969; **64**:1183–1210.
15. Monge A, Elkan C. The field-matching problem: algorithm and application. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, 1996; 267–270.
16. Jaro MA. Probabilistic linkage of large public health data files. *Statistics in Medicine* 1995; **14**:491–498 (disc: P687–P689).
17. Bilenko M, Mooney R, Cohen W, Ravikumar P, Fienberg SE. Adaptive name matching in information integration. *IEEE Intelligent Systems* 2003; **18**(5):16–23.
18. Asher J, Fienberg SE. The administrative records experiment in 2000: an application to population count estimation via triple systems estimation. *Proceedings of the Government Statistics Section*, American Statistical Association, Alexandria, VA, 2002.

19. Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society*, *Series A* (*Statistics in Society*) 2001; **164**(1):61–72.
20. Doyle P, Lane J, Theeuwes J, Zayatz L (eds). *Confidentiality*, *Disclosure and Data Access*: *Theory and Practical Applications for Statistical Agencies*. Elsevier: Amsterdam, 2001.
21. Fienberg SE. Statistical perspectives on confidentially and data access in public health. *Statistics in Medicine* 2001; **20**:1347–1357.
22. Duncan GT, Fienberg SE, Krishnan R, Padman R, Roehrig S. Disclosure limitation methods and information loss for tabular data. In *Confidentiality*, *Disclosure and Data Access*: *Theory and Practical Applications for Statistical Agencies*, Doyle P, Lane J, Theeuwes J, Zayatz L (eds). Elsevier: Amsterdam, 2001; 135–166.
23. Thompson MW. *The Killer Strain: Anthrax and a Government Exposed*. Harper Collins: New York, 2003.

