Disease Outbreak Detection System Using Syndromic Data in the Greater Washington DC Area

Michael D. Lewis, MD, MPH, Julie A. Pavlin, MD, MPH, Jay L. Mansfield, MAS, Sheilah O'Brien, MPA, Louis G. Boomsma, MSSI, Yevgeniy Elbert, MS, Patrick W. Kelley, MD, DrPH

Background: Many infectious disease outbreaks, including those caused by intentional attacks, may first present insidiously as ill-defined syndromes or unexplained deaths. While there is no substitute for the astute healthcare provider or laboratorian alerting the health department of unusual patient presentations, suspicious patterns may be apparent at the community level well before patient-level data raise an alarm.

Methods: Through centralized Department of Defense medical information systems, diagnoses based on International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes are obtained daily from 99 military emergency rooms and primary care clinics across the Washington, DC, region. Similar codes are grouped together in seven diagnostic clusters that represent related presenting signs, symptoms, and diagnoses. Daily monitoring of the data is conducted and evaluated for variation from comparable historic patterns for all seven syndrome groups. Geospatial mapping and trend analysis are performed using geographic information systems software. Data were received on a daily basis beginning in December 1999 and collection continues. The data cut-off date for this manuscript was January 2002.

Results: Demographic breakdown of military beneficiaries covered by the surveillance area reveals a broad age, gender, and geographic distribution that is generalizable to the Washington DC region. Ongoing surveillance for the previous 2 years demonstrates expected fluctuations for day-of-the-week and seasonal variations. Detection of several natural disease outbreaks are discussed as well as an analysis of retrospective data from the Centers for Disease Control and Prevention's sentinel physicians–surveillance network during the influenza season that revealed a significantly similar curve to the percentage of patients coded with a respiratory illness in this new surveillance system.

Discussion: We believe that this surveillance system can provide early detection of disease outbreaks such as influenza and possibly intentional acts. Early detection should enable officials to quickly focus limited public health resources, decrease subsequent mortality, and improve risk communication. The system is simple, flexible, and, perhaps most critical, acceptable to providers in that it puts no additional requirements on them.

Medical Subject Headings (MeSH): epidemiology, syndrome, military medicine, disease outbreaks, District of Columbia, sentinel surveillance, ambulatory care, information systems (Am J Prev Med 2002;23(3):180–186)

Background

The first objective in the 5-year strategic plan of the Department of Defense (DoD) Global Emerging Infections Surveillance and Response System (GEIS) is the improvement of surveillance systems for the

From the Division of Preventive Medicine, Walter Reed Army Institute of Research (Lewis, Kelley), Silver Spring, Maryland; Department of Defense Global Emerging Infections System (Pavlin, Mansfield, Elbert, Kelley), Silver Spring, Maryland; Vector Research, Inc. (O'Brien), Alexandria, Virginia; and U.S. Army Center for Health Promotion and Preventive Medicine (Boomsma), Aberdeen, Maryland

Address correspondence and reprint requests to: Michael D. Lewis, MD, MPH, U.S. Army Medical Component, Armed Forces Research Institute of the Medical Sciences (USAMC-AFRIMS), Bangkok, Thailand; U.S. Mail: USAMC-AFRIMS, APO AP 96546-5000 USA. E-mail: michael.lewis@thai.amedd.army.mil.

Information for the full text of this article is available via AJPM Online at www.ajpm-online.net.

early detection of emerging infections.[1] Surveillance needs encompass both laboratory-based and clinically based surveillance. Since many emerging infections, including those caused by biological terrorist attacks, may first present insidiously as ill-defined syndromes or unexplained deaths due to otherwise rare agents, rapid outbreak detection is a challenge. Increased awareness of the threat of biological warfare and terrorism emphasizes the need for timely surveillance based on symptoms as well as definitive laboratory diagnoses. Should an unannounced biological attack occur, the first sign could be an increase in healthcare utilization, probably by patients with relatively common symptomatology, such as anthrax-infected persons with respiratory complaints. If the first wave of patients is small or spread out over a large geographic area, it may not be initially obvious to individual health-

care providers, even those trained to recognize unusual disease patterns. Traditional disease surveillance has relied on an "astute clinician," a laboratory noting an unusual pathogen, or alert community members to recognize disease, confirm clinical suspicions, and notify the appropriate public health authorities.[2]

Prompt recognition of disease outbreaks and rapid diagnosis may permit the institution of prophylactic measures to decrease morbidity and mortality. Prompt recognition and rapid epidemiologic characterization may also help focus limited response resources and improve risk communication. The effects of many disease agents may be prevented or ameliorated with available antibiotics or vaccines; however, the window of opportunity is short. This underlies the need for a rapid surveillance system that does not depend on confirmed diagnoses, but rather detection of aberrant patterns. In focusing appropriate outbreak response measures and helping to identify specific etiologies, real-time surveillance based on syndromes or syndrome groups provides the quickest way to recognize and respond to many natural or unnatural disease outbreak scenarios.

Over the past decade, there has been a growing awareness concerning the nation's vulnerability to the potential use of biological pathogens for terrorism. In the summer of 1999, physicians and researchers at the Walter Reed Army Institute of Research's (WRAIR) Division of Preventive Medicine were aware of the interest of several groups in using syndrome groups for surveillance.[3] Since 1999, the New York City (NYC) Department of Health has been monitoring and evaluating emergency medical service (EMS) calls on a daily basis to identify increases in respiratory illnesses that might represent any infectious disease outbreak.[4] In the Boston, Massachusetts area, efforts by Lazarus et al.[5] to develop a system using automated medical records for analysis of lower respiratory syndrome groups were based on the groupings developed at WRAIR[5] in 1999 by Dr. Michael Lewis. Recognizing that the greater Washington DC area would be a high-value target for any terrorist attack and that no such system existed in the national capital area, we developed a syndromic surveillance system that allows for daily analysis of outpatient data from DC-area military treatment facilities (MTFs). We call this system ESSENCE—Electronic Surveillance System for the Early Notification of Community-Based Epidemics—and began receiving data on a daily basis beginning in December 1999 and have continued to the present. This system helps address a critical need that currently exists in preparedness programs for weapons of mass destruction in the United States[6] and in Washington DC in particular.

The need for further development of systems like ESSENCE or the NYC system was further brought to light during a weapons-of-mass-destruction response

exercise, code named TOPOFF. TOPOFF was conducted in May 2000 using a hypothetical attack, with the causative agent of plague in Denver, Colorado. Interviews with participants by the Johns Hopkins Center for Civilian Biodefense Studies revealed the general perception that U.S. public health systems are not prepared should such an event occur. Needs highlighted by senior health officials included a system that could deliver real-time data tabulating the number and locations of persons affected, as well as allow rapid collection and analysis of epidemiologic data.[7] While efforts have been made to address this deficiency, we believe that ESSENCE is the first large-scale system to rely on near-real time, patient-level diagnostic data. Details of the development of ESSENCE are discussed as well as some analysis of our first 2 years of experience.

Methods
Hospitals, Clinic Types, and Locations

There are 26 MTFs in a 50-mile radius of Washington DC. We identified all emergency rooms and primary care, internal medicine, pediatrics, family practice, flight medicine, and occupational health clinics at each MTF. We also included infectious disease clinics because of the unique nature of their patients. A total of 99 clinics providing primary and infectious disease care were identified for surveillance.

Capture of Daily Data

In 1997, the DoD's Ambulatory Data System (ADS) became operational in the DC area. ADS was developed for routine management purposes and is used to quantify the amount and type of work being done by military healthcare providers. For every patient encounter within the DoD, a Standardized Ambulatory Data Record (SADR) is generated and matched with patient demographic data. These data include visits by active-duty personnel and their families as well as retirees and other nonactive-duty beneficiary visits. The SADR is usually completed at the place and time that the patient encounter occurs. The provider completes the SADR with applicable known or presumed diagnoses from the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM), along with the patient's disposition, procedures, and other data. Encounters are coded even if the specific cause of illness is not evident during the visit. Most ICD-9 codes chosen reflect this initial diagnosis and may include syndrome-based codes, such as cough and fever, in addition to presumptive diagnoses, such as pneumonia or influenza.

Data were received by ESSENCE on a daily basis beginning in December 1999 and collection continues. The data cut-off date for this manuscript was January 2002. Each day, SADR forms from all MTFs in the DoD are electronically sent to a centralized system, and a direct information feed to DoD–GEIS computers allows the analysis of data from the Washington DC area clinics. Data are visualized by DoD–GEIS personnel within 1 to 3 days of the initial patient visit, depending on the timeliness of the data input by providers

Table 1. Syndrome clusters with three most commonly used International Classification of Disease, version 9 (ICD-9) codes

ICD-9 code (short description)	Count in 2000
Syndrome group 1 = coma/sudden death	
780.01 (coma)	43
798.9 (death, unattended)	7
799.1 (arrest, respiratory)	41
Total codes[a]	91
Syndrome group 2 = GI (gastrointestinal)	
558.9 (gastroenteritis, noninfct, nec)	7709
008.8 (enteritis, viral, nos)	1917
787.01 (nausea with vomiting)	2654
Total codes[a]	12,320
Syndrome group 3 = RESP (respiratory infection)	
466.0 (bronchitis, acute)	12,983
079.99 (infection, viral, nos)	23,281
462 (pharyngitis, acute)	28,538
Total codes[a]	64,802
Syndrome group 4 = neurologic	
293.0 (delirium, acute)	182
322.9 (meningitis, nos)	102
348.3 (encephalopathy, nos)	420
Total codes[a]	704
Syndrome group 5 = dermatological, infectious	
057.9 (exanthemata, viral, nos)	442
057.0 (erythema infectiosum)	65
057.8 (exanthemata, viral, nec)	49
Total codes[a]	556
Syndrome group 6 = fever	
038.8 (septicemia, nec)	8
038.9 (septicemia, nos)	453
780.6 (fever)	6597
Total codes[a]	7058
Syndrome group 7 = hemorrhagic manifestations	
287.3 (thrombocytopenia, primary)	676
287.5 (thrombocytopenia, nos)	183
782.7 (ecchymoses, spontaneous)	70
Total codes[a]	929

[a]Total codes for all ICD-9 codes for each syndrome group, including those not shown.
nos, not otherwise specified; nec, not elsewhere classified.

and the frequency of data transmission from the MTF. The requirement for GEIS to obtain these data represents no additional requirements for clinicians or clinic administrators, thus overcoming a major objection of many providers to other surveillance methodologies. It may be possible to decrease the time lag between patient visit and data arriving at the ESSENCE system through more frequent transmission of data, and efforts are being made to facilitate this.

Syndrome Groups

Since significant variability exists in diagnostic terms assigned to similar patients among providers and clinics, similar ICD-9-CM codes are grouped together in seven clusters that represent presenting signs, symptoms, and diagnoses (Table 1), which reflect common symptomatology of the most potential biological agents and naturally occurring outbreaks. An outbreak of influenza or inhalational anthrax would most likely present on a population level as an increase in respiratory complaints or fever, while an outbreak of smallpox may

first be detected through an increase in rashes and fever. To account for the large variation in reporting of many common presenting complaints and diagnoses by providers, the groups we developed encompass all possible ICD-9-CM codes for signs, symptoms, and diagnoses within each symptom group. For example, a patient presenting with a gastrointestinal complaint might be coded by one provider as having infectious enteritis (ICD-9-CM code 009.0); another provider might code the same patient as having nonspecific diarrhea (ICD-9-CM code 787.91); yet a third provider may make a presumptive diagnosis without laboratory confirmation and code the patient as having shigellosis (ICD-9-CM code 004). Any of these ICD-9-CM codes being reported would fall under the "gastrointestinal" group. Placing these similar codes into groups decreases the variability of the data and allows more accurate monitoring of background diagnostic rates.

Data Analysis and Geomapping

By establishing baseline levels of these seven groups, fluctuations in the groups can be monitored on a daily basis, and significant increases above the baseline as a result of natural or unnatural events can be detected through data analysis. Significant increases are reported immediately by telephone to public health officials at the Office of the U.S. Army Surgeon General for event monitoring and decision making on initiating an outbreak investigation. These officials also have secure website access to ESSENCE data so that they can assist with the daily monitoring, interpretation, and follow-up of aberrations in their areas of responsibility. Autoregressive analysis is performed using the SAS package, Version 8 (SAS Institute, Cary NC), which takes into consideration day-of-the-week and seasonal variation. The count of people coming in to clinics each day is intuitively related to the previous day's count and the one for the same day in the previous week. Because of occasional clinic closures due to weather and routine weekday–weekend schedules, the autocorrelation structure is often ambiguous, and that is accounted for in the statistical model.[8] Weekly trends can also be influenced by a holiday effect. For example, if Independence Day is on Monday, instead of an expected several-fold increase in patient counts from Sunday, approximately the same number of people come to the clinics as on a weekend day. In addition, in this situation there is often a significant increase in counts on Tuesday and Wednesday. Therefore, the regression part of the model is composed of two categorical variables for holiday and after-holiday effects plus dates for increased sensitivity in slow linear increases of counts. The R^2 for the syndrome groups gastrointestinal (GI), respiratory (RESP), and fever is in 0.7 to 0.9 range. The other four syndrome groups, whose daily counts are much smaller, do not show obvious weekly trends, and are not necessarily affected by holidays, are modeled using Poisson regression. Alerts are generated if the count for that day exceeds a 95% confidence interval ($p < 0.05$) of expected values (Figures 1 and 2).

Geomapping for data visualization is done with the geographic information systems (GIS) tool ArcView. Cases are plotted by patient home ZIP code (871 five-digit ZIP code locations fall within the study area). Through the use of an animated function (ArcView Tracking analyst extension), the cases can be "tracked" across the geographic region for a
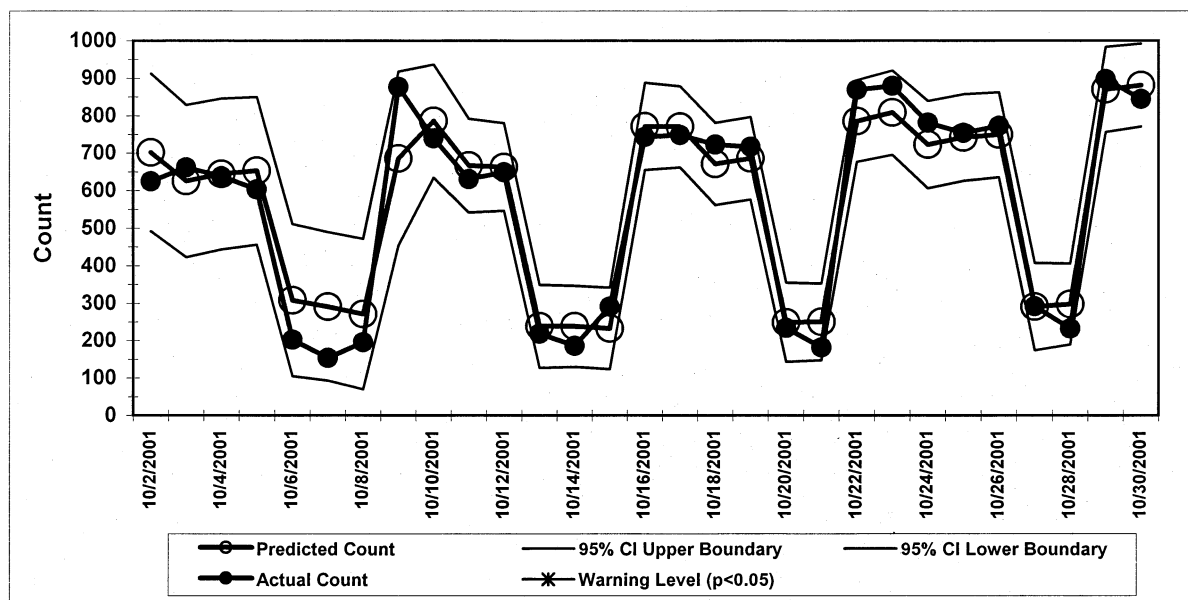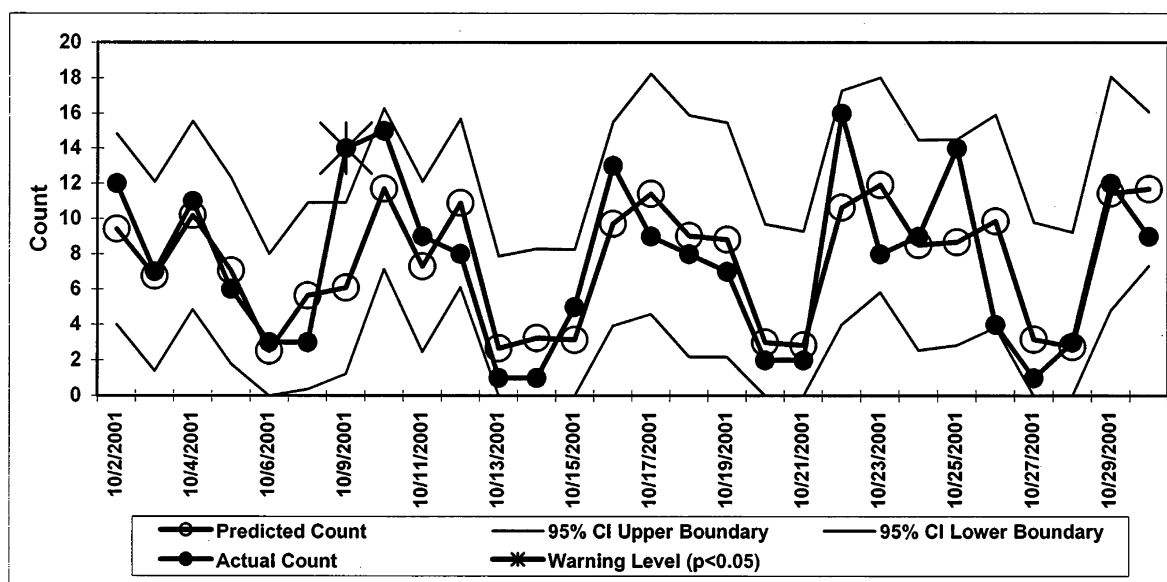
Figure 1. Actual and expected daily counts of respiratory syndrome category visits during a 1-month period in the Washington DC area. **Note:** At no time did the actual count exceed the warning level.

variable time period, portraying a "moving picture" of cases. Cases can also be aggregated by ZIP code to portray cumulative counts for each ZIP code for a requested time period. Viewing the cases through GIS allows observation of trends across the entire study area and a quick "drill down" to view local areas and individual cases. These GIS functions can help determine if a syndrome outbreak has a geographic component and could aid in locating the source of the disease outbreak if it were from a geographic point source. GIS may also help aid in predicting the extent of the affected population.

Results

Initial insights into ambulatory data patterns were derived from historic ADS data. Analysis of the demographic breakdown revealed a broad gender and age distribution similar to the general population. The geographic distribution based on home ZIP code of the population served by MTFs is also distributed throughout the Washington DC area. There are concentrated areas near larger military installations (e.g., Fort Bel-

Figure 2. Actual and expected daily counts for coma syndrome category during a 1-month period in the Washington DC area. **Note:** On October 10, the warning level was exceeded, as noted by the change in the line marker.
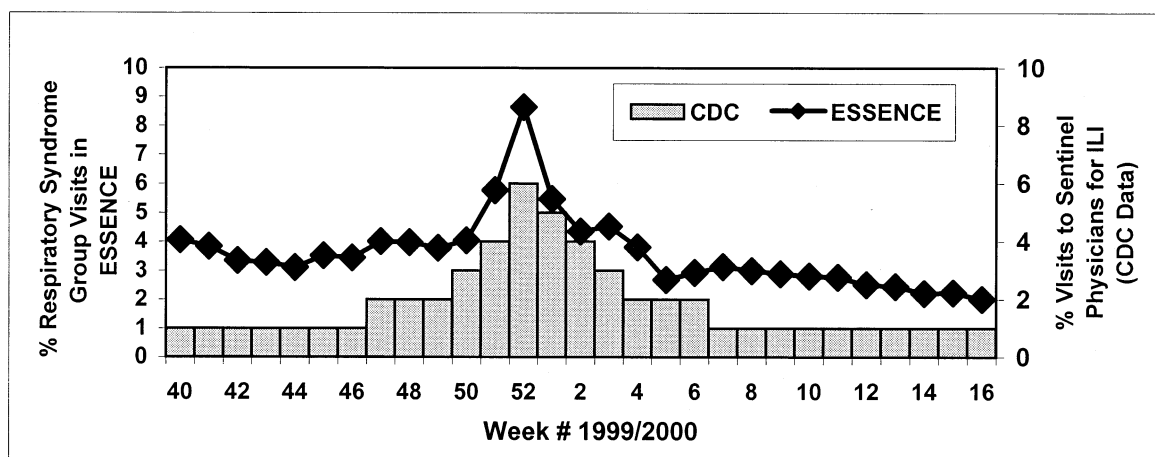
Figure 3. The percentage of outpatient visits for respiratory infections recorded in ESSENCE compared to the CDC's sentinel physician–surveillance report of percentage of visits for influenza-like illness (ILI); Spearman rank correlation coefficient = 0.704, $p < 0.001$.

voir, Virginia; and Fort Meade, Maryland), but many areas are represented throughout the 50-mile radius. We feel that the population served by MTFs in the Washington DC area is generalizable to the civilian population.

Significant day-of-the-week and seasonal differences were evident, as well as reporting variations among the clinics. These are very prominent on weekends when fewer than half the clinics are open. The modeling described previously takes into consideration these daily changes and registers only a statistically significant increase when the total counts exceed what is expected for that specific day. Using percentage of total visits for specific syndromes to determine anomalies is also possible. We compared the sensitivity of ESSENCE respiratory syndrome data with the Centers for Disease Control and Prevention (CDC) sentinel physician–surveillance network for the 1999–2000 influenza season (Figure 3).[9] Data for just the Washington DC area from the CDC during this time period is not complete enough to be used for comparison. Using the CDC data for the southeast United States, the same trend during the same time window was apparent. We calculated a Spearman rank correlation coefficient of 0.704 using the two data sources. This demonstrates a relationship between the two at a p value of <0.001.

While this represents an important historical evaluation of ESSENCE, the system was designed for near real-time recognition of potential epidemics. Within several weeks of receiving data on a daily basis on December 1, 1999, ESSENCE was able to detect a significant increase in the number of respiratory syndrome group counts. Telephonic confirmation by WRAIR researchers confirmed that MTF emergency rooms were seeing a large number of influenza cases. Visualization of the baseline for the respiratory syndrome group was compared to the previous year, and it

was quickly noted that while respiratory counts did not exceed the previous year's numbers, they were occurring 6 weeks earlier. This was later confirmed when examining the CDC influenza data.

Immediately following the events of September 11, 2001, the U.S. Army Surgeon General directed that ESSENCE begin examining DoD MTF data worldwide. In January 2002, three almost simultaneous outbreaks of diarrheal disease were detected at Fort Knox, Fort Monmouth, and the San Diego Marine Corps Recruit Depot (Figure 4). A subsequent epidemiologic investigation in San Diego revealed that a rotavirus was the causative pathogen.

Discussion

While no public health surveillance system can meet everyone's expectations, ESSENCE has definite strengths and weaknesses that are worth exploring in the context of the guidelines for evaluating surveillance systems published by the CDC.[10] ESSENCE was designed to take advantage of existing DoD patient encounter information systems and inherently will have trade-offs in system attributes. Its most positive attribute is that it is invisible to the clinician and adds no additional work not already required by the DoD. For the clinician, ESSENCE is therefore simple and acceptable, and, because it is based on extracting data from an existing system, as stable and flexible as the ADS itself.

Timeliness, data quality, sensitivity, and positive predictive value (PPV), however, present the biggest challenges for ESSENCE. The timeliness of the system, although excellent compared with traditional surveillance systems, needs to be improved. Since ADS was not intended to be used as a real-time surveillance system, rapid data transfer was not originally necessary. We are increasing the data entry and transfer speed with a goal
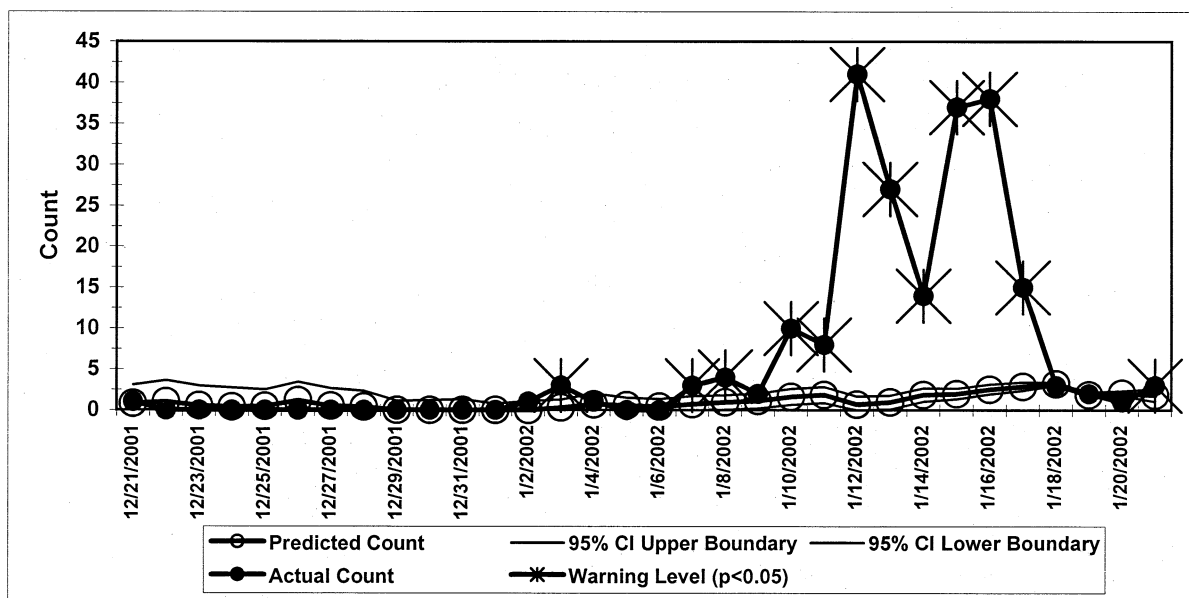
Figure 4. Gastrointestinal syndrome group counts for the San Diego Marine Corps Recruit Depot during December 2001 and January 2002. Note the dramatic increase in the number of patients seen in the second week of January.

to receive the majority of data within a 24-hour period. In the near future, the entire ADS may, in fact, become truly "real-time."

Any system that relies on the judgmental assignment of an ICD-9-CM diagnosis is going to have questionable data quality. For example, at the four-digit level, we determined that diarrhea could be reported by a clinician 174 different ways. By grouping those codes into one group, in this case GI, we are decreasing that questionable data quality as much as possible from the system's standpoint. However, it does not address the accuracy of the data actually being put into the system by the clinician in the first place. Determining the optimal sensitivity and PPV of ESSENCE or any system that relies on syndrome reporting is a challenge. A judgment must be made between timeliness and quality. Concerning the ability to detect outbreaks, which is the main reason ESSENCE was implemented, significant events may be missed if the system is not sensitive enough. Conversely, a system too sensitive will trigger multiple false alarms. In addition, because case definitions are not standard and the true incidence of a disease in an outbreak is unknown, it is difficult to estimate the PPV of ESSENCE, although across time the PPV is most likely very low. During an outbreak as the incidence of disease rises, the PPV should also rise. Determining the cut-off for sensitivity and PPV is an ongoing analytical problem that should continue to be refined as more historic data are collected. The specificity of ESSENCE, or any other syndrome-based system, will always be insufficient to detect diseases that occur in extremely low numbers, such as West Nile virus cases or even the few anthrax cases that occurred

through the U.S. Postal System in October 2002. The key piece to any system, however, is not 100% reliance on the technology to give the correct answer. There can be no substitute for human interfacing to make a good clinical judgment based on gathering all information available. In that way, the number of false alarms and unnecessary outbreak investigations can be limited. It must always be remembered that ESSENCE or any other system is a tool, not the answer in and of itself.

Integration of Other Military Information Systems

To further strengthen ESSENCE, we are currently investigating the possibility of adding several modules to the military-based, outpatient syndromic system. Pharmaceutical, laboratory, and radiologic examination data from all MTFs are being captured by DoD information systems. Accessing these data may provide a separate but complementary and confirmatory model or may be added directly to this system as variables. Tracking sentinel inpatient diagnoses such as respiratory failure or encephalitis may also help identify smaller events such as the West Nile virus.

Integration of Military and Civilian Systems

Although the military beneficiary population in the Washington DC area is distributed through most of the local community, they will not be the only ones who become ill in the event of a disease outbreak. Moreover, they may seek immediate care at the nearest civilian facility. We have proposed including data from civilian

emergency rooms and managed care organizations that would greatly increase the power of the statistical model, especially in localities where military beneficiaries are sparse. While obtaining timely and meaningful patient-record data from civilian facilities may prove problematic because of the lack of a unified data system like the DoD's, all hospital ancillary systems (laboratory, radiologic, and pharmaceutical), either civilian or military, use the Health Level 7 (HL-7) messaging standard, making it a universal means of performing surveillance activities. Many civilian healthcare providers also code outpatient visits with ICD-9 codes for billing purposes, and early work by Lazarus et al.[5] has demonstrated the ability to capture these data on a near real-time basis.

Other modules should also be considered. Autopsy and entomologic data from DoD and non-DoD sources may help explain or confirm findings of abnormality in syndrome groups. The number and types of EMS calls and calls to the Poison Control Center may serve as significant variables, as could the sales of over-the-counter medication (e.g., cough medicines, antipyretics, and antidiarrheals) and data from veterinary services and local zoos.

The DoD has recognized the importance of this concept of system integration. The Defense Advanced Research Projects Agency (DARPA) has initiated a 5-year, $24-million project to serve as a prototype system based on an earlier seedling project, which demonstrated that it is possible to identify an abnormal event several days before the medical community through evaluating medical record, grocery store, pharmacy, and school absentee databases.[11,12] The DARPA Bio-Surveillance System calls for the development of an integrated system using diverse databases from a variety of sources, such as described in the previous three paragraphs, as well as the development of sophisticated detection algorithms and models.[12]

Conclusion

We believe that ESSENCE can provide early detection of a disease outbreak and enable officials to quickly concentrate public health efforts to decrease subsequent morbidity and mortality. While there is no replacement for the astute healthcare worker, a syndromic surveillance system, if adequately sensitive as we have seen in our experience to date, can assist the healthcare system in recognizing unusual events. A robust, real-time surveillance system may be key to the detection of and prompt reaction to any infectious disease outbreak, whether from intentional or natural occurrence. Rapid recognition using a GIS component can allow documentation of the geographic extent of disease cases and provide essential information that allows prudent use of post-exposure prophylaxis and therapy. The syndromic surveillance system we have developed is a key step in closing the gap in surveillance that exists in the nation's ability to defend itself from serious disease threats.

References

1. Department of Defense Global Emerging Infections Surveillance and Response System. Addressing emerging infectious disease threats: a strategic plan for the Department of Defense. Washington, DC: Walter Reed Army Institute of Research, 1998.
2. Siegrist DW. The threat of biological attack: why concern now? Emerging Infect Dis 1999;5:505–8.
3. Kortepeter MG, Pavlin JA, Gaydos JC, et al. Surveillance at U.S. military installations for bioterrorist and emerging infectious disease threats. Mil Med 2000;165:ii–iii.
4. Lackelsberg J, Layton M. New York City Department of Health Alert #6 (October 5, 2001). Available at: http://nyc.gov/html/doh/html/cd/wtcg.html. Accessed March 30, 2002.
5. Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health 2001;1:9.
6. Waeckerle JF. Domestic preparedness for events involving weapons of mass destruction. JAMA 2000;283:252–4.
7. Inglesby T, Grossman R, O'Toole T. A plague on your city: observations from TOPOFF. Biodefense Q 2000;2:1–10.
8. Box, GEP, Jenkins GM, Reinsel GC. Time series analysis: forecasting and control. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1994.
9. Centers for Disease Control and Prevention. Update: influenza activity—United States and worldwide, 1999–2000 seasons, and composition of the 2000–01 influenza vaccine. MMWR Morb Mort Wkly Rep 2000;49:375–81.
10. Centers for Disease Control and Prevention. Updated guidelines for evaluating public health surveillance systems. MMWR Morb Mort Wkly Rep 2001;50:1–35.
11. Brown D. Catching the bug before it kills: Pentagon would mine data to fight bio-terrorism. Available at: www.zdnet.com/zdnn/stories/news/0,4586,2672554,00.html. Accessed January 30, 2001.
12. Defense Advanced Research Projects Agency. Bio-Surveillance System BAA Proposer Information Pamphlet (PIP). Available at: www.darpa.mil/iso/BIOS/BAA01-17.htm. Accessed January 30, 2001.