SANDIA REPORT

SAND2003-8154
Unlimited Release
Printed April 2003

Enabling Analytical and Modeling Tools for Enhanced Disease Surveillance

D.K. Manley, M.W. Koch, S.A. McKenna, R.L. Bilisoly, M.W. Trahan, A.L. Sobel, D.D. Djordjevich, M.E. Goldsby, E.L. Hoffman, M.M. Johnson, T. Nguyen, T.J. Sa, H.R. Ammerlahn, M.F. Hawley, W.B. Wilcox, A.S. Yoshimura

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2003-8154
Unlimited Release
Printed April 2003

Enabling Analytical and Modeling Tools for Enhanced Disease Surveillance

Mark W. Koch
Signal and Image Processing Systems Department

Sean A. McKenna and Roger L. Bilisoly
Geohydrology Department

Michael W. Trahan
System Technologies Department

Annette L. Sobel
Systems Analysis Department

Dawn K. Manley, Donna D. Djordjevich, Mike E. Goldsby, Edward L. Hoffman, Michael M. Johnson, Tammy Nguyen, and Timothy J. Sa
Systems Studies Department

Heidi R. Ammerlahn, Marilyn F. Hawley, William B. Wilcox, and Ann S. Yoshimura
Systems Research Department

Sandia National Laboratories
P.O. Box 969
Livermore, CA 94551-0969

ABSTRACT

Early detection, identification, and warning are essential to minimize casualties from a biological attack. For covert attacks, sick people are likely to provide the first indication of an attack. An enhanced medical surveillance system that synthesizes distributed health indicator information and rapidly analyzes the information can dramatically increase the number of lives saved.

Current surveillance methods to detect both biological attacks and natural outbreaks are hindered by factors such as distributed ownership of information, incompatible data storage and analysis programs, and patient privacy concerns. Moreover, because data are not widely shared, few data mining algorithms have been tested on and applied to diverse health indicator data.

This project addressed both integration of multiple data sources and development and integration of analytical tools for rapid detection of disease outbreaks. As a first prototype, we developed an application to query and display distributed patient records. This application incorporated need-to-know access control and incorporated data from standard commercial databases. We developed and tested two different algorithms for outbreak recognition. The first is a pattern recognition technique that searches for space-time data clusters that may signal a disease outbreak. The second is a genetic algorithm to design and train neural networks (GANN) that we applied toward disease forecasting. We tested these algorithms against influenza, respiratory illness, and Dengue Fever data.

Through this LDRD in combination with other internal funding, we delivered a distributed simulation capability to synthesize disparate information and models for earlier recognition and improved decision-making in the event of a biological attack. The architecture incorporates user feedback and control so that a user's decision inputs can impact the scenario outcome as well as integrated security and role-based access-control for communicating between distributed data and analytical tools. This work included construction of interfaces to various commercial database products and to one of the data analysis algorithms developed through this LDRD.

CONTENTS

FIGURES

TABLES

Outbreak Recognition

Data Sources

We obtained several health indicator databases for testing the outbreak recognition techniques. The first were publicly available weekly morbidity reports collected by the French Sentinel Disease Network (http://www.b3e.jussieu.fr:80/sentiweb/en/). These data contain weekly morbidity for eight diseases including influenza-like-illness. The second were obtained through partnerships with health and military agencies. California's Office of Statewide Health Planning and Development (OSHPD) provided patient-level hospital discharge data refined by date and zip code for all California hospitals from 1990-1999. Each record includes the primary diagnosis code (ICD-9-CM International Classification of Diseases, 9th Revision, Clinical Modification) as well as the major diagnostic category (MDC) and diagnosis related group (DRG) code. We also obtained monthly Dengue Fever and weather data from Bangkok, Thailand (1966-1994) and Yogyakarta, Indonesia (1985-1998).

Space-Time Clustering

We developed new space-time clustering algorithms to detect anomalies in morbidity data. Space-time clustering searches for statistically significant clusters of adverse health events in space and time, can indicate exposure to infectious diseases or localized exposure to toxins, and helps pinpoint the location of the contaminant's source. Space-time anomalies may arise not only from intentional outbreaks, but also from new and emerging infectious diseases or epidemics. This allows one to test against real historical data. Popular space-time statistical tests such as Knox (Knox 1964a and 1964b), Mantel (Mantel 1967), and *k*-Nearest Neighbor (Jacquez 1996) use point-type data. These tests look for correlations between cases in time and cases in space, and also use distribution-free or nonparametric techniques to determine the statistical significance of a cluster. See (Williams 1984) for an excellent review of space-time clustering tests of individual case histories.

Nonparametric Space-Time Clustering

Non-parametric space-time clustering algorithms include those of Knox (1964b) and Mantel (1967) and a newly developed technique, the Poisson Kolmogorov-Smirnov (K-S) test (McKenna 2002). This new technique looks for anomalous increases in the incidence of disease and then examines the spatial clustering of these occurrences with a two-dimensional K-S test. The Knox test determines the sum of all cases that occur within a user-specified time and distance threshold. A large sum relative to the reference distribution indicates a significant cluster of disease occurrence. The reference distribution is based on sums calculated on the same number of cases in the same locations, but with randomized times. The Mantel test is similar to the Knox, but no user-defined space and time thresholds are required. The Mantel test examines the raw sum of the products of the spatial and temporal distances between disease occurrences. These two tests were modified to work with near-real time aggregated data by considering the amount by which a user-specified threshold number of cases was exceeded in any region to represent a single "case" and then calculating the values of each test and the distribution of reference sums within a user-specified moving time window.

We developed an additional technique to determine the amount of temporal and spatial clustering using the 2-D K-S test. This Poisson K-S test is actually semi-parametric in that the determination of anomalous increases in the incidence of a disease is based on the assumption of a Poisson distribution for the number of diseases during any time period. Anomalously high increases in the number of diseases in any spatial region are determined as incidence that exceeds a user-specified probability level in the Poisson distribution (e.g., 0.999), where the intensity parameter of the Poisson distribution is defined using the previous N time steps. Clustering of the spatial coordinates of these anomalously high time steps is examined with a 2-D K-S test. The 2-D K-S test was developed for

work in astronomy and determines whether or not two groups of data that are randomly scattered across a 2-D domain are clustered relative to one another. For application to disease data, the spatial locations of regions with anomalous increases in disease incidence are compared to the same locations of regions without anomalous increases. If the 2-D K-S test indicates a high probability of clustering, then the start of an epidemic is signaled.

These techniques were applied to a simulated data set and to the OSHPD data set at the zip code (1640 zip codes) and daily time step resolution for four different ICD-9-CM diagnoses (influenza, unspecified asthma, sore throat and salmonella). Results show that the Knox test is more sensitive to increased numbers of cases and thus produces a higher false positive rate than does the Mantel test. The Mantel produces good results in the detection of the onset of smaller, localized epidemics. An advantage of the Mantel test over the Knox test is that user-defined values of the time and distance thresholds do not have to be set. A drawback of both of these tests is that the significance of the test must be determined from a reference distribution that is obtained through multiple calculations of the test statistic. For large data sets, these calculations can be computationally expensive.

Parametric Space-Time Clustering

Most space-time clustering approaches require individual patient data. To protect the patient's privacy, we have extended these approaches to aggregated data and have embedded this extension in a sequential probability ratio test (SPRT) framework. Instead of looking for correlations between cases in space and cases in time (Knox 1964a and 1964b, Mantel 1967, and Jacquez 1996), we take a more direct approach. Here we take an analysis date and for every region search back in time and over nearby regions for all possible space-time clusters. The user determines the maximum cluster size in time and space, and a binomial SPRT determines the cluster's significance. This approach not only indicates the presence of a cluster, but also its location and size.

The real-time and sequential nature of health data makes the SPRT an ideal candidate. The SPRT keeps gathering and combining observations as long as the statistical test has a value between the upper stopping boundary A, and the lower stopping boundary B. Once the test goes above A or below B, the SPRT cluster-detector makes a decision. These upper and lower stopping boundaries determine the desired false alarm (FA) error rate and the desired missed-detection (MD) error rate. We have extended the SPRT to handle the spatial and temporal dependencies often found in space-time data. The result of space-time clustering gives the statistical significance of a cluster at every location in the surveillance area and can be thought of as a "health-index" of the people living in this area.

Medical surveillance data are far from Gaussian and often best described as a Levy process, which is fractal with large tails, infinite moments, and non-closed form density functions. Data resolved by date and zip code may be sparse with zero or very few events per location. These reasons forced us to develop statistical tests for which very little is assumed about the distribution underlying the disease incidence. In the SPRT test, we compare disease incidence to a critical threshold. The critical threshold gives the value of incidence for when the background likelihood equals the anomaly likelihood. Below the critical threshold the incidence of adverse-health events more likely belongs to the background and above the critical threshold the incidence of adverse-health events more likely belongs to the anomaly (e.g. epidemic or outbreak) class. Using the critical threshold we can transform the distribution of any stochastic process into a binomial distribution.

We tested our approach with the French morbidity data for influenza-like illness and the OSHPD data for influenza. For both databases, we show that space-time clustering can detect a flu epidemic up to 21 to 28 days earlier than a conventional periodic regression technique (Koch 2002). We have also tested using simulated anthrax attack data overlaid upon the OSHPD data for respiratory illness as the background incidence of disease. Results show we do very well at detecting an attack as early as the

second or third day of when infected people start becoming severely symptomatic (Koch 2002). These results are discussed further in Section 2.2.1.

Neural Networks

In addition to developing outbreak detection techniques, we also developed an algorithm for disease forecasting. We used a genetic algorithm to design and train neural networks (GANN). The GANN is trained on a subset of data and then predicts the incidence of disease and epidemics.

We examined monthly values for temperature, saturation deficit, precipitation, relative humidity, and cases of Dengue Fever. For Bangkok, we also had population data and forecasted the disease incidence. This technique predicted the disease incidence thirty, sixty, and ninety days in advance (Trahan 2001). For Yogyakarta, we had average sea surface temperature data in addition to temperature, saturation deficit, precipitation, relative humidity, and cases of Dengue Fever. Unlike the Bangkok data, the Yogyakarta data did not include population size making it impossible to forecast the incidence of disease. Instead, we chose to forecast whether there would be an epidemic, arbitrarily defined as 100 or more cases of Dengue Fever in a month, at various times in the future. The GANN predicted epidemics thirty and ninety days in advance for the Yogyakarta data (Trahan 2002).

We also attempted to apply GANN to forecast outbreaks of respiratory disease in the San Francisco Bay Area of California using the OSHPD data covering the period from January 1995 through June 1999. The data sets were very large (in excess of two million records per year for 1995-1998 and in excess of one million records for 1999), but most of the data was not useful in this effort. The data set provided a record for each patient admitted to the hospital. Each record included values for the level of care, sex, ethnicity, race, patient's county of residence, source of admission, type of admission, disposition of patient, and diagnostic codes. The only values of interest for this work were gender, county of residence (within or outside the Bay Area), major diagnostic code (respiratory illness or other), admission date, and age category. The data was aggregated by day, by week, and by month. We tried to forecast the number of cases of respiratory illness in the Bay Area within a given time frame (day, week, or month) some period of time in the future without success. We believe the failure was due to the lack of information from which the neural networks had to learn and generalize.

Forecasting outbreaks of respiratory illness is very different from forecasting outbreaks of Dengue Fever. Dengue Fever is a non-contagious, vector-borne disease. Respiratory illness, however, has many causes and in many cases is contagious. It is possible to estimate the prevalence of the Dengue Fever in the environment by measuring environmental conditions that affect the vector (the mosquito). Since respiratory illness covers a number of different diseases, there is no easy way to measure its prevalence in the environment. In the case of the Dengue Fever, the neural networks were able to learn what environmental factors had an impact on the disease. With the limited data set, the neural networks were not able to learn any such correspondences for the respiratory illness.

Distributed Information Integration

A more complete picture of potential disease outbreaks and application of outbreak recognition algorithms require integration of data from a variety of distributed sources. For example, patient records, syndromic data, ambulance calls, and environmental data may be owned and maintained by several different organizations and may be stored in incompatible databases and programs. Patient privacy concerns may further limit the sharing of information. We developed interfaces to enable communication between a variety of databases and models that allowed applications to draw upon information from distributed sources. This framework integrates security and access control so that participants can maintain complete control and ownership of their information. Each participant determines who can access their information and the level of detail they can access.

Medical Surveillance Application

As a first prototype, we developed a medical surveillance application to query and display distributed patient records. This application incorporated need-to-know access control and incorporated data from standard commercial databases including SQL Server and Microsoft Access. As shown on the following pages, in Figure 1 the user is presented with a query panel that contains the data fields for which she has access based on her login role. Once the user formulates a query, the application queries the distributed databases and presents the result set as illustrated in Figure 2.

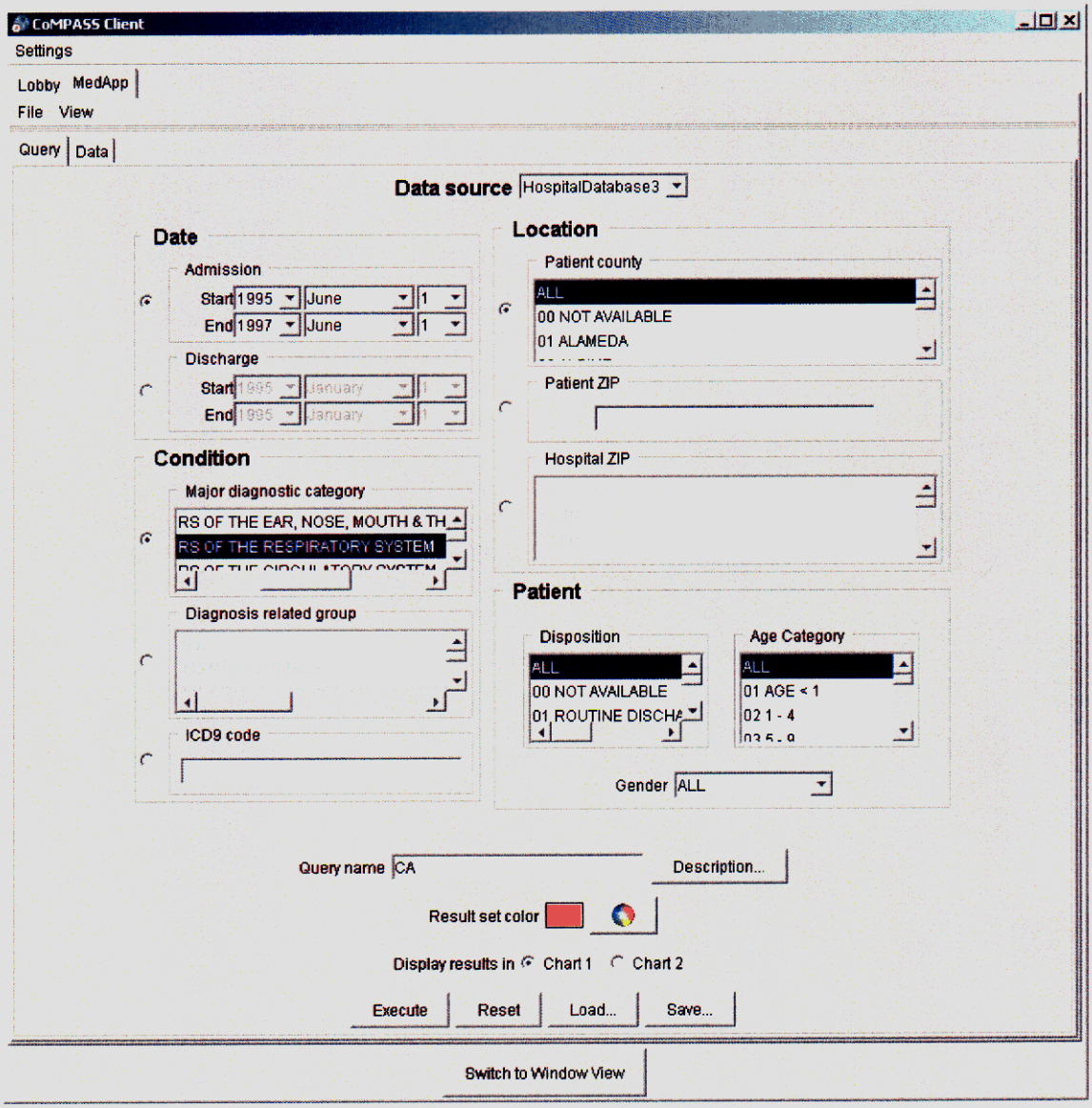

Figure 1. Query input panel from distributed application.

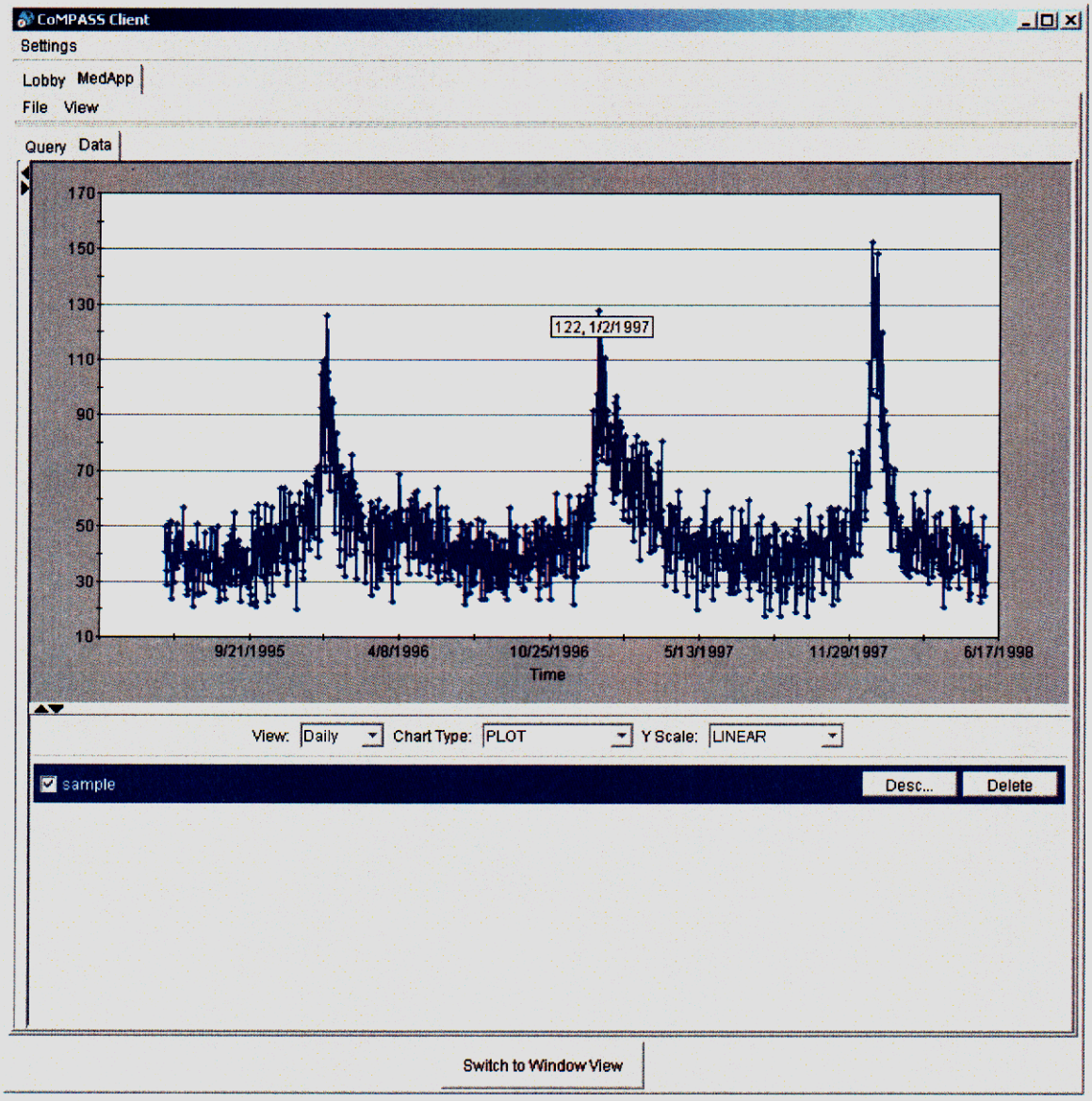

Figure 2. Query result panel from distributed application.

Weapons of Mass Destruction Decision Analysis Center

In combination with other internal funding, we delivered a distributed simulation capability, the Weapons of Mass Destruction Decision Analysis Center (WMDDAC), to synthesize distributed information and models for earlier recognition and improved decision-making in bioterrorism events. The architecture includes user feedback and control so that decision inputs can impact the simulated scenario as well as integrated security and role-based access control.

The first version of WMDDAC was used to simulate an anthrax release in the San Francisco Bay Area. The user plays the role of a public health officer and is able to make decisions during the simulation based on information representative of that which a real public health officer would access during an epidemic. Background health data includes hospital data from the OSHPD database. The

user can conduct a simple epidemiological investigation by examining various public health reports such as morbidity and death reports, as well as implement a prophylaxis strategy by activating a simulated National Pharmaceutical Stockpile and prophylaxis distributions centers and by directing people to seek treatment.

Integration of Space-Time Clustering Algorithm

A common request during early demonstrations was for analytical tools to help understand whether the reported health data was normal or indicative of an epidemic. Because a tool like the space-time clustering algorithm could aid a decision maker in the early detection of an anthrax attack, we integrated the algorithm discussed in Section 1.2.2 into the simulation. This work included construction of user interfaces and module development to enable the algorithm to communicate in the simulation framework. Sample output from a simulated release is shown in Figure 3.

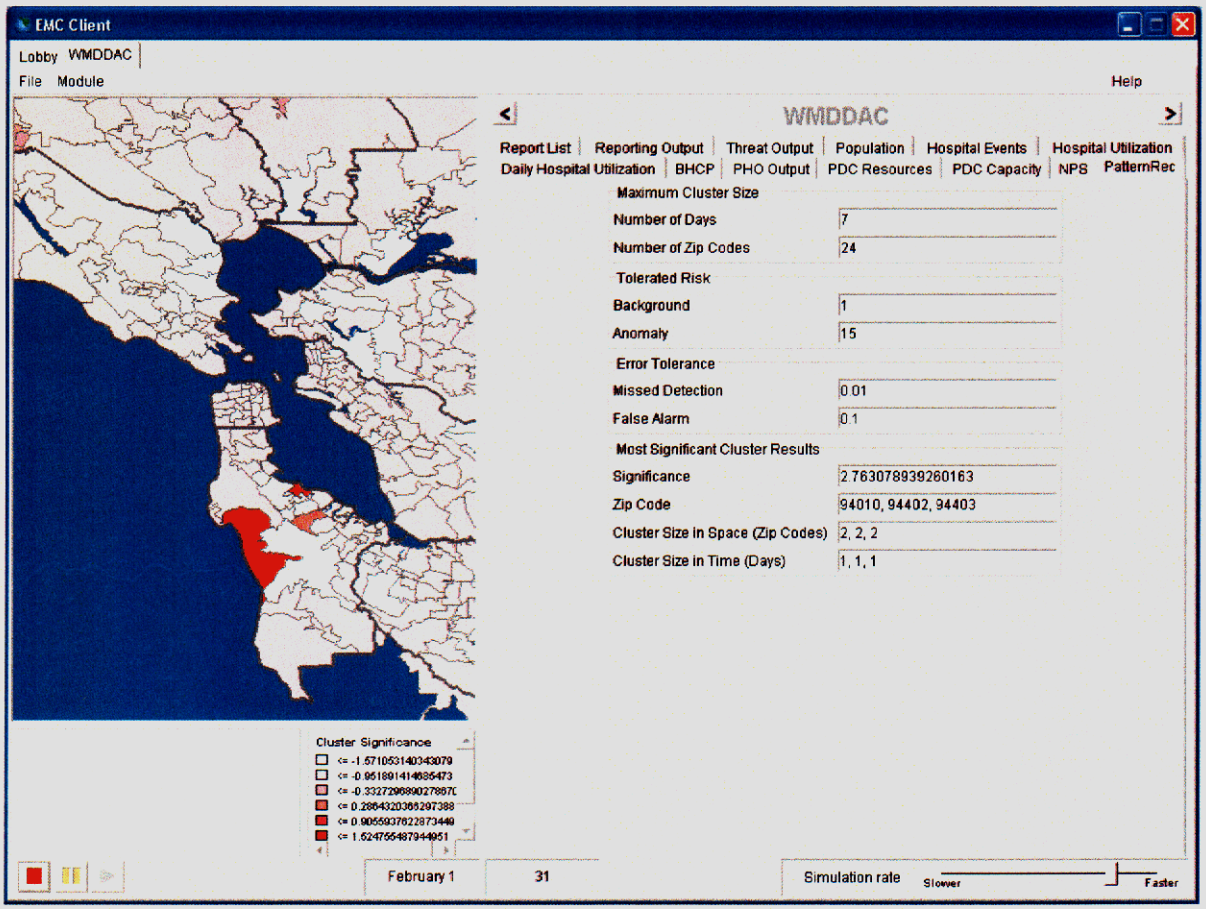

Figure 3. Sample output from parametric space-time clustering algorithm model incorporated into WMDDAC simulation.

In its early stages, a health care professional might mistake anthrax for another respiratory illness. Figure 4 shows the incidence of diseases and disorders of the respiratory system for the years 1991 through 1997 in the Bay Area counties of Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma. Respiratory illness has a seasonal component with peaks in the winter months and also a slightly increasing trend due mostly likely to the increasing population.

These data provide training data and the simulated anthrax release gives a signal to add to the background.

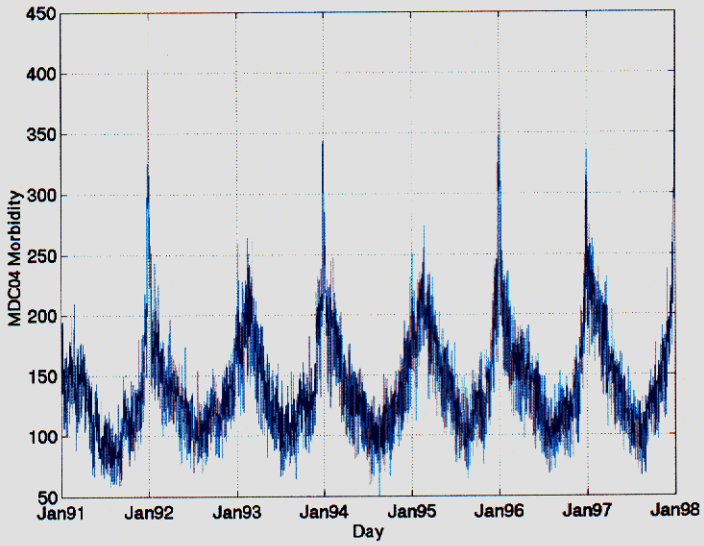

Figure 4. Incidence of respiratory illness from the OSHPD database.

Table. 1. Number of Simulated Anthrax Cases for each Plume versus the Days after the Start of the Attack.

Number of Cases		Days After Simulated Anthrax Attack									
		0	1	2	3	4	5	6	7	8	9
Plume #	1	0	0	0	2	13	32	54	57	42	39
	2	0	0	0	1	36	98	129	118	118	102
	3	0	0	0	2	20	54	72	83	70	40
	4	0	0	0	1	0	2	3	0	2	2
	5	0	0	0	0	5	25	39	28	32	19
	6	0	0	0	1	3	19	14	15	10	18
	7	0	0	0	0	18	48	47	52	32	48
	8	0	0	0	0	3	9	15	24	24	9
	9	0	0	0	0	12	38	75	61	60	41
	10	0	0	0	0	0	4	2	3	4	2
	11	0	0	0	0	2	13	13	12	9	5
	12	0	0	0	0	3	7	11	12	6	6
	13	0	0	0	1	26	64	76	69	51	60

Table 1 shows the number of simulated anthrax cases versus the days after the start of the attack for thirteen different release scenarios. The table shows that in most cases infected people do not become severely symptomatic until at least the third or fourth day after the attack. Simulated plumes 4 and 10 produce very few cases due to the small attack size and their location in sparsely populated areas.

These attacks disappear into the background. Dark gray areas indicate when the space-time clustering algorithm detects a cluster with no false alarms and greater than 99% probability of detection. The lighter gray areas indicate detection, but at a reduced probability of detection (above 90%). The table shows we do very well at detecting an attack by the second or third day of when infected people start becoming severely symptomatic.

Summary

An enhanced medical surveillance system will require three components: the collection of appropriate data, a framework for distributed information integration, and analytical tools to sift through the information. This project has addressed the latter two components through development of an information unification framework, space-time clustering algorithms, and genetic algorithm designed and trained neural networks.

We developed space-time clustering algorithms and applied these to outbreak detection in morbidity data. For influenza epidemics, this technique detected epidemics up to 3-4 weeks earlier than conventional periodic regression techniques. We also tested against a simulated anthrax attack overlaid upon background respiratory illness data. Our algorithm consistently detected the attack within 2-3 days after infected people became symptomatic.

In addition to detection algorithms, we developed a genetic algorithm to design and train neural networks for forecasting disease incidence. This method worked well for predicting vector-borne, non-contagious disease when trained with disease incidence and environmental parameters that influence the vector. This method was also trained with patient records for respiratory illness, but found no correlation between the patient record parameters and disease. The algorithm requires more information on parameters that influence the disease.

We developed a framework for distributed information integration with integrated security and access-control. We built two applications upon this framework. The first is an application to query and display patient records from distributed data sources. The query input and result set reflected the user's access privileges to the data. The second is a decision analysis capability, the Weapons of Mass Destruction Decision Analysis Center (WMDDAC), which integrates distributed data and models for simulated bioterrorism events. The user is able to make decisions during the simulation that impact the outcome of the scenario. Because a tool like the space-time clustering algorithm could aid a decision maker in the early detection of an anthrax attack, we integrated the algorithm into the simulation. This work included construction of user interfaces and module development to enable the algorithm to communicate in the simulation framework.

References

Jacquez, G. M., 1996. "A k nearest neighbor test for space-time interaction," Statistics in Medicine, vol. 15, pp. 1935-1949.

Koch, M. W., McKenna, S. A., Bilisoly, R. L., "Syndromic Surveillance using Parametric Space-Time Clustering" Sandia National Laboratories Report SAND2002-3747, 2002.

Knox, G., 1964a. "Epidemiology of childhood leukemia in Northumberland and Durham," British Journal of Preventive and Social Medicine, vol. 17, pp. 17-24.

Knox, G., 1964b. "The detection of space-time interactions," Applied Statistics, vol. 13, pp. 25-29.

Mantel, N., 1967. "The detection disease clustering and a generalized regression approach," Cancer Research, 27, No. 2, pp. 209-220.

McKenna, S. A., Bilisoly, R. L., Koch, M. W., "Syndrome Surveillance using Nonparametric Space-Time Clustering," Sandia National Laboratories Report, 2002.

Trahan, M. W., Sobel, A. L., "A Novel Technique for Forecasting Outbreaks of Vector-Borne Diseases," Sandia National Laboratories Report SAND2002-0605C, 2002.

Trahan, M. W., Sobel, A. L., "Transnational and Homeland Defense: Software Tools for Biological Threat Reduction," BTR 2001: Unified Science & Technology for Biological Threat Reduction, 2001.

Williams, G. W., 1984. "Time-space clustering of disease," in Statistical Methods for Cancer Studies, R. G. Cornell, ed., Marcel Dekker Inc., pp. 167-227.

Distribution

1	MS 0188	Donna Chavez, 1011
1	MS 0735	Roger Bilisoly, 6115
1	MS 0735	Sean McKenna, 6115
1	MS 0844	Mark Koch, 15352
1	MS 1203	Alan Zelicoff, 5320
1	MS 1207	Michael Trahan, 5914
1	MS 1219	Annette Sobel 5907
1	MS 9004	John Vitko, 8100
1	MS 9201	Ann Yoshimura, 8112
10	MS 9201	Dawn Manley, 8114
1	MS 9201	Donna Djordjevich, 8114
1	MS 9201	Edward Hoffman, 8114
1	MS 9201	Heidi Ammerlahn, 8112
1	MS 9201	Larry Brandt, 8112
1	MS 9201	Marilyn Hawley, 8112
1	MS 9201	Michael Johnson, 8114
1	MS 9201	Mike Goldsby, 8114
1	MS 9201	Pat Falcone, 8114
1	MS 9201	Timothy Sa, 8114
1	MS 9201	William Wilcox, 8112
1	MS 9951	Duane Lindner, 8101
1	MS 9951	Len Napolitano, 8100
3	MS 9018	Central Technical File, 8945-1
1	MS 0899	Technical Library, 9616
1	MS 9021	Classification Office, 8511/Technical Library, MS 0899, 9616 DOE/OSTI via URL