Chapter 6

SYSTEM ASSESSMENT AND EVALUATION

Knowing how systems perform under various scenarios is important. We need to examine with which level of sensitivity and how quickly they can detect an outbreak or recognize a bioterrorism attack. Knowing the error rate of a system can help make decisions regarding how much effort is needed to investigate an alarm. The performance of the algorithms for outbreak characterization determines the amount of information they provide (e.g., sets of affected individuals, the outbreak size, and disease spreading rate), which provide important input for response planning.

Substantial costs can be incurred when developing or managing syndromic surveillance systems and investigating possible outbreaks based on the outputs of these systems (Reingold, 2003). For example, as reported in (Doroshenko et al., 2005), the annual cost of the NHS Direct Syndromic Surveillance System is about $280,000 and the usefulness of surveillance systems for early detection and response is yet to be established. Assessing the performance of surveillance systems is of significant importance for improving the efficacy of the investment in system development and management (Buehler et al., 2004).

As we discussed Chapter 4, dozens of different data analytical methods have been developed in the literature, and each method has its own limitations and strengths in different circumstances. One algorithm might work better when the size of the outbreak infected population is in a particular range. Another algorithm might have the lowest error rate in a slow-building but not a sudden-surge outbreak. Most researchers agree that no single algorithm can effectively cover the wide spectrum of all possible situations (Aamodt et al., 2006; Siegrist et al., 2004). As such, thoroughly evaluating different systems

H. Chen et al., *Infectious Disease Informatics: Syndromic Surveillance for Public Health and BioDefense*, Integrated Series in Information Systems 21, DOI 10.1007/978-1-4419-1278-7_6,

and analytical methods can provide important clues about their strengths and weaknesses, and their applicability in various application scenarios.

However, there fundamental difficulties in the evaluation of outbreak detection methods. The difficulties involve specification of the aberration of interest, and determining whether the aberration is of public health importance, caused by an infectious disease outbreak or not. In short, outbreaks are difficult to define precisely. Measurement of the validity of an outbreak detection method can be very complicated.

In this chapter, we first present a system evaluation framework that outlines three linked pieces of work evaluating communication components, outbreak detection algorithms, and system interface features. We then focus on evaluating outbreak detection algorithms along with syndrome classification algorithms. We then discuss the evaluation of data collection and information dissemination components and the system interface features. For each evaluation task, we introduce the commonly used measurement metrics. We also report representative evaluation results from a number of system evaluation studies employing the discussed measures.

1. SYNDROMIC SURVEILLANCE SYSTEM EVALUATION FRAMEWORK

CDC's Guidelines for Evaluating Surveillance Systems aim to address "the need for (a) the integration of surveillance and health information systems, (b) the establishment of data standards, (c) the electronic exchange of health data, and (d) changes in the objectives of public health surveillance to facilitate the response of public health to emerging health threats (e.g., new diseases)" (Buehler et al., 2004).

Many existing evaluation studies follow the guidelines of CDC's evaluation framework. This evaluation framework consists of a series of steps requiring the involvement of stakeholders, the description of system components, and the gathering of credible evidence regarding the system performance. It can serve as a checklist to guide the design and implementation of an evaluation procedure. Along with the description of the step-by-step tasks, relevant standards are also provided for each of the tasks for assessing the quality of the evaluation activities. Simplicity, flexibility, data quality, acceptability, sensitivity, predictive value positive (PVP), representativeness, timeliness, and stability need to quantified or described. These standards will be further developed later in this chapter when we discuss evaluation of specific components of syndromic surveillance systems.

Our evaluation framework in general follows the CDC evaluation framework but treats major system components separately for the purpose of performance analysis, considering their differences in terms of performance metrics, and visibility to different set of users. The specific evaluation tasks include evaluation of outbreak detection algorithms, data collection and information dissemination components, and system interface features.

2. EVALUATION OF OUTBREAK DETECTION ALGORITHMS

2.1 Evaluation Methodology

Simulation is one of the well-developed computational methodologies that can be applied to testing outbreak detection algorithms' validity and reliability. Different types of simulated signals, different days of duration, and different case distributions need to be specified in a simulation study, representing a realization of the system dynamic behavior. Tunable replications of simulation also enable the examination of alternative solutions. In addition, because of its flexibility and direct mapping to real-world entities, simulation can be used for training purposes and produce useful animated visual outputs.

On the basis of the extent of data authenticity, three types of simulation are possible. One is to use real data collected from real outbreaks. However, because the number of real outbreaks is small (Siegrist and Pavlin, 2004), it is very difficult to test outbreak detection algorithms using completely authentic data. Simulated outbreaks can also be superimposed on real data to provide additional tests for model validity. There are fully synthetic data-based simulation and semisynthetic data-based simulation. Without actual outbreak data, simulation-based evaluation, in particular, the fully synthetic data-based simulation, often demonstrates only limited validity (Kleinman et al., 2005b).

2.2 Real Data Testing

Running outbreak detection algorithms on real data provides the strongest and most direct validity tests. But the lack of surveillance data with real disease outbreaks makes it difficult for real data testing. There are very few published evaluation works that use real data with sufficient sample size to test outbreak detection algorithms. These few studies include the retrospective analysis by Hogan et al. (2003), a retrospective evaluation study (Ivanov et al., 2003), and the Bio-ALIRT Biosurveillance Detection Algorithm Evaluation program (Siegrist and Pavlin, 2004).

In Hogan et al.'s study, two types of real data – sales of electrolyte products and hospital diagnoses – were collected from six urban regions in three states for the period 1998 through 2001. The gold standard outbreaks are 18 significant increases in respiratory and diarrheal disease in the data. Time gain using the sales of electrolyte products to signal outbreaks of respiratory and diarrheal diseases in children compared with the hospital diagnoses were seen.

Ivanov et al. (2003) conducted a retrospective evaluation study evaluating chief complaints and the EWMA detection algorithm employing gold standard outbreaks obtained from a dataset derived from the Utah Hospital Discharge Database for the years 1998–2001 inclusive.

In the Bio-ALIRT Biosurveillance Detection Algorithm Evaluation program conducted by Siegrist et al., real historic deidentified data were obtained from five metropolitan areas over 23 months. Two natural disease outbreak cases in the data identified and labeled by an outbreak detection group were used as the gold standard. The study reports the difficulty in determining how quickly an algorithm might detect an attack is due to the fact that minimal data exists for an actual biologic attack. The limitations of real data testing are discussed, including the uncertainty about the exact start date and size of outbreaks and the inability to examine algorithm outbreak-detection capabilities under a substantial number of diverse conditions.

2.3 Fully Synthetic Data Testing

To address the data problem, synthetic data or semisynthetic data are often used in characterizing the performance of the outbreak detection algorithms.

Simulators are designed to generate the surveillance data such as illness incidences, drug purchases, physician visits that can best mimic the realization with careful characterization of an outbreak event and sick people's healthcare seeking behaviors.

A number of methods have been applied to generate these synthetic data. One is to use the outbreak detection algorithm itself by running it backwards to generate the illness incidence data. This kind of evaluation process was used to evaluate WSARE (Wagner et al., 2006).

Another method composes the shapes of outbreak signals by looking at the historical outbreaks. Figure 6-1 shows five temporal distributions used in one simulation study (Jackson et al., 2007). The temporal distributions are extracted from the epidemic curves of historic outbreaks, representing several ways in which a pathogen could spread through a community. They then specify the range of outbreak signal durations, and ranges of sizes of populations affected to generate a number of simulated outbreaks.

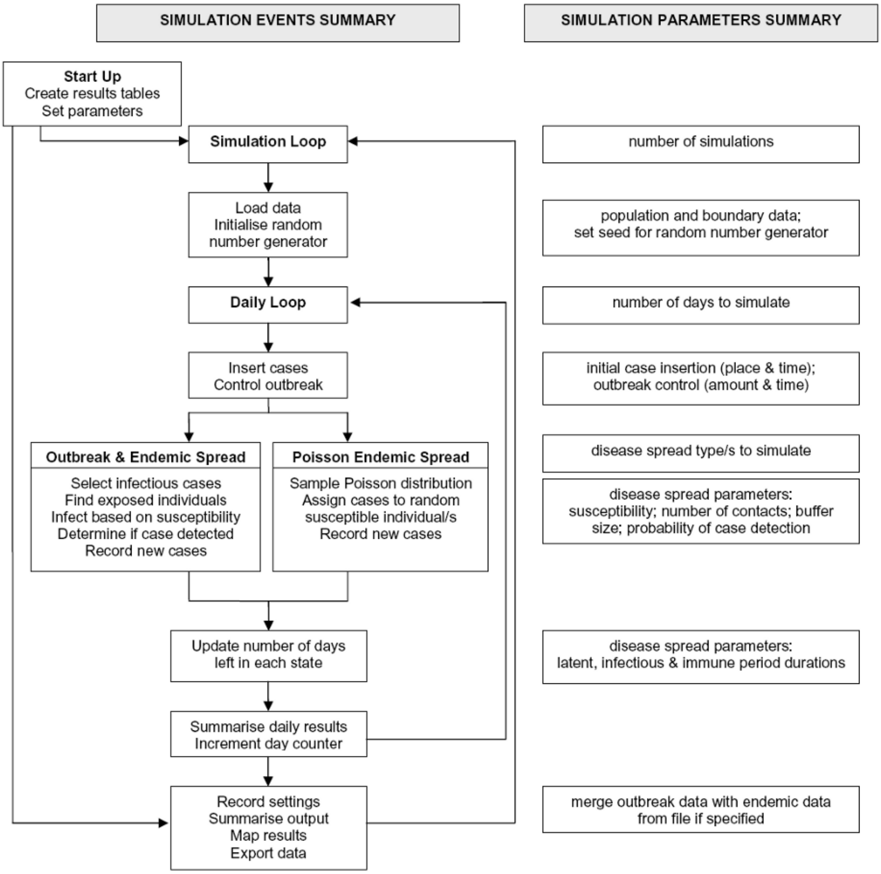

SIMULATION EVENTS SUMMARY
SIMULATION PARAMETERS SUMMARY
Start Up
Create results tables
Set parameters
Simulation Loop
Load data
Initialise random number generator
Daily Loop
Insert cases
Control outbreak
Outbreak & Endemic Spread
Select infectious cases
Find exposed individuals
Infect based on susceptibility
Determine if case detected
Record new cases
Poisson Endemic Spread
Sample Poisson distribution
Assign cases to random susceptible individual/s
Record new cases
Update number of days left in each state
Summarise daily results
Increment day counter
Record settings
Summarise output
Map results
Export data
number of simulations
population and boundary data; set seed for random number generator
number of days to simulate
initial case insertion (place & time); outbreak control (amount & time)
disease spread type/s to simulate
disease spread parameters: susceptibility; number of contacts; buffer size; probability of case detection
disease spread parameters: latent, infectious & immune period durations
merge outbreak data with endemic data from file if specified

Figure 6-1. Simulation process diagram (Watkins et al., 2007).

A third method is agent-based method. Agent-based simulators (e.g., BioWar (Carley et al., 2003)) are also used to generate the surveillance data that best represent the realistic outbreak events by modeling the social and epidemiological characterization of public health status, which describes how people acquire diseases, manifest symptoms, seek information, and seek care. RODS also developed a CityBN (City Bayesian Network) simulator to validate the WSARE algorithm. The CityBN simulator runs on a large Bayesian network whose structure and parameters are created by hand. The Bayesian network introduces temporal dynamics based on a variety of factors such as weather and food conditions (Wong et al., 2005).

Researchers also proposed to apply state-transition modeling techniques to simulate disease outbreaks (Watkins et al., 2007). The spread of infectious diseases transmitted by person-to-person contact in daily time steps can be modeled (the process diagram is shown in Figure 6-1). The model parameters are specified as disease-specific infectivity and susceptibility at individual level based on the SEIR (Susceptible, Exposed, Infectious, Recovered) approach that is commonly used to describe the epidemiology of infectious diseases. The software was developed using the MapBasic programming language for the MapInfo Professional GIS environment.

The fully synthetic data-based testing is advantageous because of the data availability and control over the evaluation process. The size of the outbreak, the spatial distribution, and many other characteristics can be changed to simulate variable outbreak events. Precise information about outbreaks can be used to measure the effectiveness of the methods under testing objectively and precisely. However, the synthetically generated data usually embody many assumptions to match the evaluated algorithms' assumptions, thus possessing limited validity. Typically, the use of the synthetic data testing is restricted to early stage testing of algorithms.

2.4 Semisynthetic Data Testing

An alternative method to generate surveillance testing data takes the approach of adding simulated outbreak cases to the real data streams. This approach is sometimes referred to as "injecting" or "spiking" events into real surveillance data collected during nonoutbreak periods (Wagner et al., 2006). More sophisticated injection techniques model the outbreaks with the shape and noise level derived from surveillance data collected during real outbreaks. The high-fidelity detectability experiments (HiFIDE) are available for noncommercial use.

Most of the evaluation studies take this approach for system evaluation (Reis et al., 2003; Goldenberg et al., 2002). In the evaluation work of EARS (Hutwagner et al., 2005a), for instance, 56,000 sets of artificially generated case-count data are generated based on 56 sets of parameters using a negative binomial distribution with superimposed outbreaks. The ESSENCE II system is evaluated using simulated bioterrorism events with estimated patterns from the literature (Lombardo et al., 2003).

The semisynthetic approach provides greater validity than the fully synthetic data-based testing. It allows for flexible manipulation of outbreak sizes and the shapes of the spikes as well as the time courses of each injected event. In-depth understanding of the dynamics of real outbreaks is crucial for the fidelity of the injected outbreaks.

2.5 Evaluation Metrics for Outbreak Detection Algorithms

The main concerns regarding anomaly detection algorithms include how significant the signal needs to be to trigger an alarm, how early an outbreak can be detected, and how reliable the alarms are. Various aspects of outbreak detection algorithms need to be evaluated using different evaluation criteria. Such criteria include the quantification of sensitivity, predictive value positive, timeliness, false alarm rate, generalized ROC curves, and average run length. These criteria are in line with the CDC evaluation guidelines (CDC, 2001) and the prior literature (Buehler et al., 2004; Romaguera et al., 2000). Table 6-1 summarized the outbreak detection metrics in the most commonly used representations in the literature. A more detailed summary of detection algorithm evaluation metrics can also be found in (Buckeridge et al., 2005b).

Three metrics – sensitivity, false alarm rate or the alternative measure to the false alarm rate – predictive value positive and timeliness, are most commonly seen in the literature (Buckeridge et al., 2004; Sonesson and Bock, 2003). Sensitivity measures the probability that an alarm is correctly triggered when an outbreak indeed occurs. False alarm rate measures the

Table 6-1. Outbreak detection metrics.

Terms	Descriptions
Sensitivity	The proportion of outbreaks that an algorithm detected correctly (Wagner et al., 2006)
Specificity	The proportion of nonoutbreaks days without alarms (Wagner et al., 2006)
Predictive Value Positive (PVP)	The proportion of alarms signaled as outbreaks are truly outbreaks (CDC, 2001)
Timeliness (time-to-detection)	The difference between the date of the first true alarm and a reference date (e.g., a date established as a start date of an outbreak by expert consensus) (Wagner et al., 2006)
False alarm rate	The proportion of nonoutbreak time periods (days or weeks depending on the organization of the time series) on which an algorithm signals alarms (Wagner et al., 2006)
ROC curve	Plot of sensitivity versus false alarm rate
AMOC curve	Plot of timeliness against false alarm rate
ARL^0	Expected run length until the first false alarm (Sonesson and Bock, 2003)
ARL^1	Expected run length until an alarm (Sonesson and Bock, 2003)

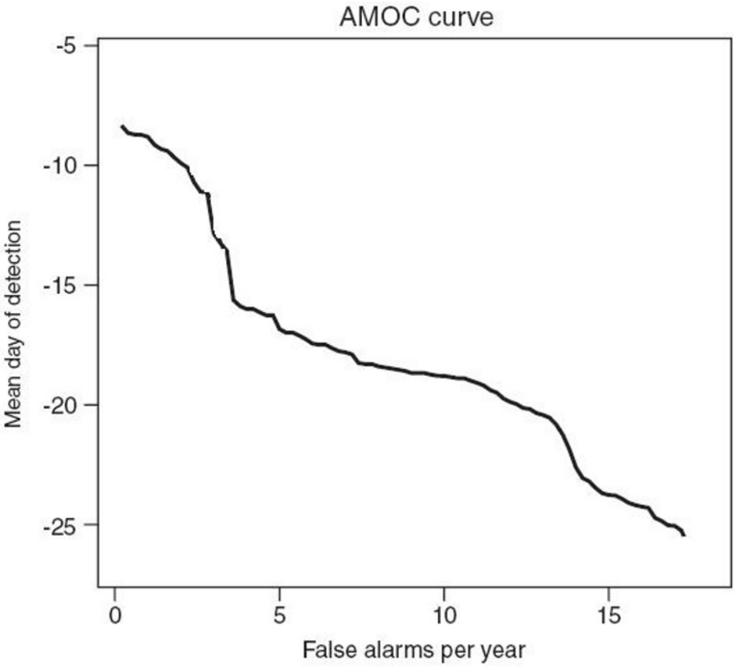
ROC: Receiver Operating Characteristic

AMOC: Activity Monitoring Operating Characteristic

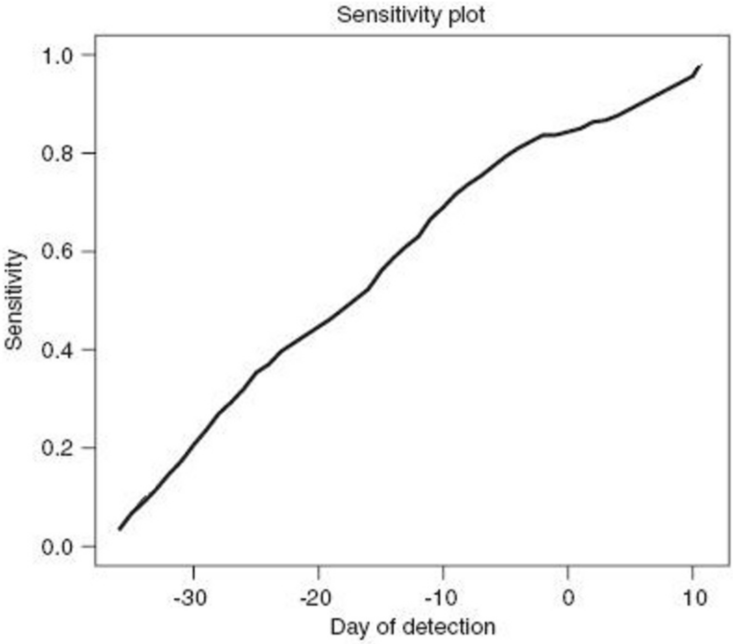
ARL: Average Run Length

probability that an alarm is triggered when there is no outbreak. The measurement of sensitivity and PVP for a syndromic surveillance system is often complicated by the absence of an appropriate gold standard (German, 2000). A gold standard is assumed to be accurate and can be used to validate the signals produced by an outbreak detection system.

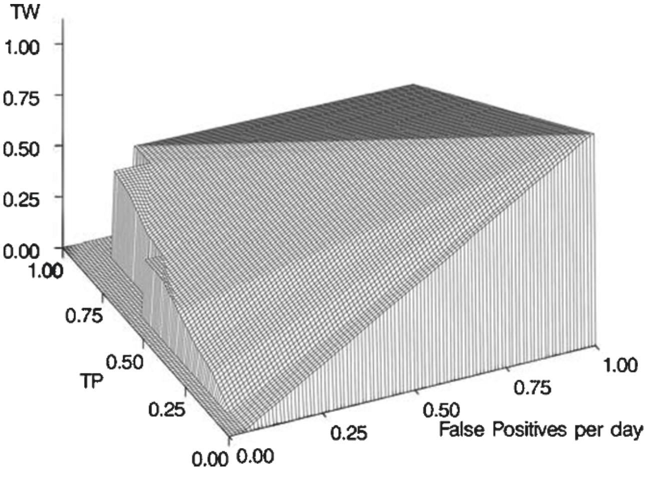
There exists a tradeoff when trying to achieve good performance among multiple evaluation criteria (Buckeridge et al., 2004; Siegrist and Pavlin, 2004). The Receiver Operating Characteristics (ROC) curve and the area beneath it are further evaluation metrics that plot sensitivity against false alarm rate (Reis and Mandl, 2003). Through the AMOC (Activity Monitoring Operating Characteristic) curve plotting timeliness against false alarm rate, the evaluators can easily read the tradeoff between the false alarm rate and the timeliness.



(a) Fictional AMOC curve.



(b) Timeliness-ROC curve.



(c) Timeliness-ROC surface for surveillance assessment; vertical dimension is the timeliness weight (Kleinman and Abrams 2006).

Figure 6-2. Fictional AMOC curve, timeliness-ROC curve and timeliness-ROC surface.

Figure 6-2a and b show a fictional AMOC curve and timeliness-ROC curve, respectively. A three-dimensional generalized ROC curve is proposed by Kleinman and Abrams (2006). The 3-D ROC curve incorporates the time of detection and produces the timeliness-ROC surfaces (an example is shown in Figure 6-2c). By incorporating sensitivity, specificity and timeliness into single metrics, the proposed approach simplifies the comparison of different methods' performance.

Timeliness measures the proportion of time gained by an early detection algorithm compared with a reference signal (e.g., the clinical diagnosis of an anthrax case). As a means to measure the efficiency of detection algorithms, it refers to how fast an aberration is signaled. The expected delay time can be denoted by $ED(t) = E\left[\max(0, t_{\mathrm{A}} - t) \middle| \tau = t\right]$, where the time of change is $\tau = t$, and the time of alarm triggering is t_{A}. However, the timeliness of a surveillance process should also include the delay in the process data collection and case reporting in addition to the time for disease case identification. The timeliness of the data collection process is now generally indicated by the frequency of data uploading, either manually or automatically, by the data providers. A real time surveillance system must feature a real time and automated data collection mechanism as discussed in Chapter 3.

2.6 Summary of Representative Evaluation Studies

We have conducted a systematic review of syndromic surveillance system performance evaluation studies. Out of 55 publications that claim to evaluate syndromic surveillance systems, 32 reported evaluation results or system experiences with varying degrees of detail. Two systems were compared with a reference detection system. Timeliness versus sensitivity plotting was provided in 19 quantitative evaluations of algorithms' detection performance (e.g., WSARE, SaTScan, and RSVC). Twelve systems reported sensitivity and false alarm rate through the ROC curve. A few evaluations such as the BioALIRT evaluation program are conducted to examine the algorithms from different systems for side-by-side comparison.

For a selected set of detection algorithms, we provide details about evaluation design and settings (e.g., the data sets used, the outbreak detection methods evaluated, and the simulated outbreak patterns). We also present the evaluation results according to the performance metrics used in the evaluation. However, as the simulation models and datasets used for evaluating each algorithm differ, a conclusive performance report is not feasible.

Table 6-2. Summary of evaluation results on a selected set of syndromic surveillance systems.

Syndromes	Dataset	Evaluated system	Evaluated methods	Outbreak patterns	Criteria	Results
Respiratory infection	Daily hospital ED visit data (Mar. 2002–Dec. 2003) from Hillsborough County, Florida	EARS	P-chart C2 C3 MA [EARS] EWMA	Slow-building or sudden-surge trend	CARL ROC	The use of C2 and P-chart for timely surveillance is suggested when the syndromic data are moderately correlated (Zhu et al., 2005)
National and state pneumonia, influenza data and hospital influenza-like illness	56,000 sets of artificially generated case-count data based on 56 sets of parameters, with superimposed outbreaks	EARS (Hutwagner et al., 2005a)	C1-mild, C2-medium, C3-Ultra, the historical limits method and the seasonally adjusted CUSUM	Log normal, a rapidly increasing outbreak; inverted log normal, a slowly starting outbreak; a single-day spike	Sensitivity, specificity, time to detection	These simulations demonstrate that the methods for aberration detection that require little baseline data, C1, C2, and C3, are as sensitive and specific as the historical limits and seasonally adjusted CUSUM methods
Six syndromes (unspecified)	Simulated data generated from surveillance data in ED during several large public events in the United States	EARS (Hutwagner et al., 2005b)	C1-mild, C2-medium, and C3-ultra	The aberrations were added to the baseline data using a random binomial distribution	Levels of sensitivity, specificity, false positive rates	For the six syndromes, sensitivity for C1, C2, and C3-models averaged 48, 51, and 54%. The specificities averaged 98, 98, and 96%, respectively. The average false-positive rates were 32, 29, and 42%, respectively
Respiratory and gastrointestinal syndromes	Historic de-identified data obtained from five metropolitan areas over 23 months	2003 Bio-ALIRT algorithm evaluation	SPC, Bayesian change-point, wavelet methods, RODS, ESSENCE, EARS, and General Dynamics and IBM (Siegrist et al., 2004)	Actual outbreaks embedded in the data	Timeliness and sensitivity versus false-positive rates	The best algorithms (anonymous) were able to detect all of the outbreaks at false-alert rates of one every 2–6 weeks. However, whether certain algorithms were better overall than others was not determined

Syndromes	Dataset	Evaluated system	Evaluated methods	Outbreak patterns	Criteria	Results
Simulated data	Hospital ED respiratory syndrome counts, office visits, respiratory counts, OTC influenza medication sales, and school absentee totals.	ESSENCE II	Methods in ESSENCE II	Simulated bioterrorism events with estimated patterns from the literature	Sensitivity, specificity	The number of infected people is varied to achieve a detector performance with a sensitivity of 0.95 and a specificity of 0.97 (Lombardo et al., 2003)
Influenza	ICD-coded chief complaints (Dec 4, 1999 – Dec 1, 2000)	RODS	Serfling method [RODS]		Sensitivity, PVP, timeliness	For a one-year period, the detectors had sensitivity of 100% and PVP of 50% for RS and 25% for IS. The timeliness of detection using ICD-9–coded chief complaints was one week earlier than the detection using Pneumonia and Influenza deaths (Espino and Wagner, 2001; Tsui et al., 2001)
Influenza-like illness (ILI)	New York City Emergency Medical Services (EMS) ambulance dispatch data	New York City EMS	Data quality Case detection algorithm		Sensitivity, PVP	The selected call types had a sensitivity of 58% for clinical ILI, and a PVP of 22% (Greenko et al., 2003)

3. EVALUATION OF DATA COLLECTION AND INFORMATION DISSEMINATION COMPONENTS

The system components for data collection and information dissemination need to be evaluated in terms of HIPAA compliance, scalability, and flexibility.

HIPAA privacy rules govern the obligations and reporting requirements of healthcare data (CDC, 2003). HIPAA security regulations require methods that protect data from disclosure in transport. To be HIPAA compliant, data collection and dissemination components of syndromic surveillance systems need to provide security measures such as data encryption, secure sockets, secure shell tunneling, or the use of a virtual private network.

System scalability and flexibility indicate how scalable a syndromic surveillance system is in monitoring new diseases, accommodating new syndrome categories, or incorporating new types of data. Geographic coverage should be able to be expanded with small costs as additional healthcare facilities and jurisdictions participate. In addition, systems that use standard data formats (e.g., in electronic data interchange) can easily interoperate with other systems and thus might be considered more flexible and more scalable (CDC, 2001).

4. ASSESSMENT OF INTERFACE FEATURES AND SYSTEM USABILITY

4.1 System Usability Evaluation Methodology

To complete our discussion of system evaluation, the performance of operational systems bringing in the users' operation experiences need to be evaluated. The effectiveness (or value) of a syndromic surveillance system depends greatly on the outcome associated with their use of the system. The evaluation process usually employs two methodologies: controlled experiment and field testing. Controlled experiments consider the users' experience with the interaction with the system interfaces for completion of a particular operation. Field testing evaluates operational systems mainly for the measurement of the benefit, and the cost from a perspective of societal utility (Wagner et al., 2006). It takes into account how long it takes to deploy a system, what the system failure rate is, and so on.

4.2 System Usability Evaluation Metrics

In the evaluation work for the BioPortal system, Hu et al. (2005) applied a number quantitative or qualitative metrics for system usability evaluation. (1) Task accuracy: the correctness of the user generated analysis results using the system referenced to the experts' analysis results; (2) Task efficiency: measuring the amount of time a person needs to complete an analysis task; (3) User satisfaction: end-user satisfaction typically encompasses system content, accuracy, output format, use, and timeliness; (4) Perceived usefulness: it refers to the extent to which a person considers a system useful in his or her work role and has been shown to affect user adoption significantly; (5) Perceived ease of use: the ease of use of a system, as perceived by individual users refers to the degree to which a person believes that using a particular system will be free of effort.

Wagner summarized a group of measurable system benefits and cost related system features in his recent work on field testing of biosurveillance systems (Wagner et al., 2006). The metrics are: (1) Benefits from expected reductions in mortality and morbidity through earlier detection; (2) Benefits [usefulness, simplicity, representativeness (CDC, 2001)] from expected reductions in operational costs owing to policy improvements and workflow efficiency; (3) Costs to build or purchase and install, and costs of staff time on alarms monitoring and investigation and certain other metrics.

4.3 Summary of System Usability Evaluation Studies

The evaluation study conducted by Hu et al. (2005) is representative of research examining syndromic surveillance system usability issues, such as readability, learning curve, and decision making assistance. They used the User Interaction Satisfaction (QUIS) instrument by Chin et al. (1988) to evaluate the usability of the BioPortal system, based on the Object-Action Interface model developed by Shneiderman (1998). They examined the overall reactions to the system, the screen layout and sequence, the system's capability, the terminology/information used, and subjects' ease of learning, based on a 9-Point Likert scale (Hu et al., 2005).

From a user's perspective, all relevant data must be seamlessly integrated to support the surveillance and analysis tasks that are critical to the prevention of and alerts about particular disease events or devastating outbreaks. Data visualization support is also critical; the value of a syndromic surveillance system is greatly affected by the extent to which the system can present data and analysis results in an easily comprehensible, cognitively efficient manner. Ultimately, a syndromic surveillance system must facilitate and enhance the analysis tasks by public health professionals in terms of accuracy and time requirements, using their own heuristics and preferred analysis methods.

5. SUMMARY AND DISCUSSION

Evaluation of syndromic surveillance systems is confounded by a number of factors. First, few real-world datasets are available for evaluation and comparison purposes due to the low frequency or absence of outbreaks of most diseases. Second, timeliness of detection is closely related to the timing of patient visits or medication purchases, determined by individual patients' behavior. Third, data quality and availability are seldom considered in algorithm evaluations. Incomplete data from various healthcare participants can potentially impair algorithms' detection power.

Fourth, the criteria for optimized detection performance may vary for different illnesses. Different bioterrorism agents display different temporal and spatial patterns. Botulism and toxic shock syndrome are readily detected in relatively smaller clusters, whereas detection of SARS presents a greater challenge as the syndrome is relatively less specific and the impact may be more widely spread. The incubation time and the time between exposure and symptom onset could be longer or shorter depending on the type of biologic agent. The detection power of the algorithms for rare diseases (e.g., botulism-like illness or smallpox) is yet to be reported.

Lastly, the ability of an algorithm to identify the geographic location of an outbreak was rarely measured and reported. In spatial context, the signal extent is not usually considered. For example, in a scan-like method, the radius of a detected cluster could indicate a kind of accuracy of the detection method. The cluster validity measurement techniques discussed in a few works (Halkidi et al., 2002) seem ready to check the clustering algorithms' performance.