Chapter 26

INFECTIOUS DISEASE INFORMATICS AND SYNDROMIC SURVEILLANCE

Daniel Zeng[1], Hsinchun Chen[1], and Ping Yan[1]

[1]*Management Information Systems Department, University of Arizona, Tucson, Arizona,U.S.A.*

CHAPTER OVERVIEW

Infectious disease informatics (IDI) is an emerging field that studies data collection, sharing, modeling, and management issues in the domain of infectious diseases. This chapter discusses various technical components of IDI research from an information technology perspective. Syndromic surveillance is used to illustrate these components of IDI research, as it is a widely-adopted approach to detecting and responding to public health and bioterrorism events. Two case studies involving real-world applications and research prototypes are presented to illustrate the application context and relevant system design and data modeling issues.

1. INTRODUCTION

In highly-mobile modern societies, infectious diseases, either naturally occurring or caused by bioterrorism attacks, can spread at a rapid rate, resulting in potentially significant loss of life and economic assets (Pinner, Rebmann et al. 2003; Zeng, Chen et al. 2004; Zeng, Chen et al. 2005).

Information systems are increasingly playing a significant role in developing an effective approach to prevent, detect, respond to, and manage infectious disease outbreaks of plants, animals, and humans (Buehler, Hopkins et al. 2004; Zeng, Chen et al. 2005). A large amount of data related to infectious disease detection and response is being collected by various laboratories, health care providers, and government agencies at local, state, national, and international levels (Pinner, Rebmann et al. 2003). A number of information access, analysis, and reporting systems have also been developed and adopted in various public health and homeland security application contexts. For example, in its role as the key agency responsible for human reportable diseases in the U.S., the U.S. Centers for Disease Control and Prevention (CDC) has developed computerized reporting systems for local and state health departments and is also playing a central role in coordinating standardization efforts with an aim for interoperable messaging and data/system integration. Similarly, the U.S. Department of Agriculture (USDA) is enhancing data systems for certain animal diseases (e.g., mad cow disease and foot-and-mouth disease), and the U.S. Geological Survey (USGS), through its National Wildlife Health Center (NWHC) and numerous partners, manages databases for wildlife diseases.

A significant portion of recent work, from the perspectives of both academic research and real-world system implementation, on public health and bioterrorism event detection and response, has been focused on syndromic surveillance (Lawson and Kleinman 2005; Wagner, Moore et al. 2006). Syndromic surveillance is defined as an ongoing, systematic collection, analysis, and interpretation of "syndrome"-specific data for early detection of public health aberrations. The rationale behind syndromic surveillance lies in the fact that specific diseases of interest can be monitored by syndromic presentations that can be shown in a timely manner such as nurse calls, medication purchases, and school or work absenteeism. In addition to early detection and reporting of monitored diseases, syndromic surveillance also provides a rich data repository and highly active communication system for situation awareness and event characterization.

This chapter introduces infectious disease informatics (IDI), an emerging subfield of biomedical informatics that systematically studies information management and analysis issues in the domain of infectious diseases (Zeng, Chen et al. 2005; Hu, Zeng et al. 2007). The objective of IDI research can be

summarized as the development of the science and technologies needed for collecting, sharing, reporting, analyzing, and visualizing infectious disease data and for providing data and decision-making support for infectious disease prevention, detection, and management. IDI research directly benefits public health agencies in their infectious disease surveillance activities at all levels of government and in the international context. It also has important applications in national security concerning potential bioterrorism attacks (Siegrist 1999).

This chapter emphasizes various technical components of IDI research with detailed discussions on the design and various system components of an infectious disease information infrastructure. Considering the importance of syndromic surveillance as a widely-adopted approach to detect and respond to public health and bioterrorism events, we use syndromic surveillance systems as the representative IDI application to frame the discussion.

The rest of the chapter is organized as follows. Section 2 provides a brief overview of IDI, discussing its overall objectives and the key motivating technical and policy-related challenges. In Section 2, we also present the basic technical components of an IDI system and related design considerations from an information systems perspective. Section 3 presents an introduction to syndromic surveillance systems, covering issues ranging from applicable data sources, related data analysis and modeling work, to data visualization. In Section 4 we use two real-world IDI projects to provide the readers with a concrete sense of what IDI systems look like and how they are used in specific application contexts. A major purpose of this section is to illustrate in an integrated manner how the technical issues discussed in Sections 2 and 3 are treated in applications. We conclude in Section 5 by summarizing the chapter and pointing out several ongoing trends in IDI research and application.

2. INFECTIOUS DISEASE INFORMATICS AND ITS MAJOR TECHNICAL COMPONENTS

2.1 Objectives and Challenges

IDI is an emerging field of study that systematically examines information management and analysis issues related to infectious diseases (Zeng, Chen et al. 2004; Hu, Zeng et al. 2007). The specific technical objectives of IDI research can be summarized as the development and evaluation of frameworks, techniques, and systems needed for collecting, sharing, reporting, analyzing, and visualizing infectious disease data and for providing data and decision-making support for human, animal, and plant

infectious disease prevention, detection, and management. IDI research is also concerned with studying technology adoption issues to promote real-world application of these IDI frameworks, techniques, and systems.

IDI research is inherently interdisciplinary, drawing expertise from a number of fields including but not limited to various branches of information technologies such as information integration, knowledge representation, data sharing, Geographic Information Systems (GIS), data mining, text mining, and visualization; and other fields such as biostatistics, bioinformatics, dynamical systems, operations research, and management information systems. It also has a critical policy component dealing with issues such as data ownership and access control, intra- and inter-agency collaboration, and data privacy and data confidentiality. Because of its broad coverage, IDI research and practice rely on broad participation and partnership between various academic disciplines, public health and other disease surveillance, management, diagnostic, or research agencies at all levels, law enforcement and national security agencies, and related international organizations and government branches.

The basic functions required of any IDI system are data entry, storage, and query, typically implemented as computerized record management systems deployed by local public health agencies or at local hospitals and laboratory facilities. To enable information sharing and reporting across record management systems maintained at different sites, system designers and operators need to agree on a common technical approach. This technical approach should include data sharing protocols based on interoperable standards and a Web-enabled distributed data store infrastructure to allow easy access. It also needs to provide a scalable and effective reporting and alerting mechanism across organizational boundaries and provide important geocoding and GIS-based visualizations to facilitate infectious disease data analysis.

To maximize the potential payoff of an IDI system, advanced information management and data analysis capabilities need to be made available to the users. Such information management capabilities include visualization support to facilitate understanding and summarization of large amounts of data. An important aspect of data analysis is concerned with outbreak detection and prediction in the form of spatio-temporal data analysis and surveillance. New “privacy-conscious” data mining techniques also need to be developed to better protect privacy and patient confidentiality (Kargupta, Liu et al. 2003; Wylie and Mineau 2003; Ohno-Machado, PS et al. 2004).

From a policy perspective, there are mainly four sets of issues that need to be studied and related guidelines developed (Zeng, Chen et al. 2004). The first set is concerned with legal issues. There exist many laws, regulations,

and agreements governing data collection, data confidentiality and reporting, which directly impact the design and operations of IDI systems. The second set is mainly related to data ownership and access control issues. The key questions are: Who are the owner(s) of a particular dataset and derivative data? Who is allowed to input, access, aggregate, or distribute data? The third set concerns data dissemination and alerting: What alerts should be sent to whom under what circumstances? The policy governing data dissemination and alerting needs to be made jointly by organizations across jurisdictions and has to carefully balance the needs for information and possibility of information overflow. The fourth set is concerned with data sharing and possible incentive mechanisms. To facilitate fruitful sharing of infectious disease data on an ongoing basis, all contributing parties need to have proper incentives and benefit from the collaboration.

To summarize, from an application standpoint, the ideal IDI system would include a field deployable electronic collection instrument that could be synchronized with server based information systems in public health departments (Zeng, Chen et al. 2005). Biological specimen processing would be handled in laboratory information systems that were integrated completely with epidemiological and demographic information collected from the field and with the electronically submitted data from non-public health clinical laboratory information systems. The integrated laboratory, demographic, and epidemiological information would be available for statistical and GIS analyses in real time as the data are collected. The data collected would be available to authorized users of a system that would protect identifying information of any individuals using role based user access and permissions. Data could be shared across public health jurisdictions and between public health and nonpublic health agencies where such sharing was appropriate and where only appropriate data was provided. The ideal data system would use standards for metadata, terminologies, messaging formats, and security to maintain true semantic interoperability. Data analysis, altering, and decision support would be integrated into the data stream for data validation, message routing, and data de-duplication.

2.2 Basic Technical Components and Design Considerations

This section summarizes basic technical components of an IDI system needed to provide essential data support. More advanced functionalities of IDI systems tend to be application-specific; some of these functionalities and related technological support will be discussed in Section 3.

The following technical considerations are critical to the design of a basic IDI system: data standards, system architecture and messaging standards, and data ingest and access control. In this section, we briefly discuss them in turn at the conceptual level. In Section 3, these issues will be revisited in the specific context of syndromic surveillance.

Data Standards. Data standards enable interoperability between information systems involved in disease reporting and surveillance. Data standards are also critical to provide unambiguous meaning to data and form the foundation that enables data aggregation as well as data mining. Many data standards have been developed in health care and public health informatics, causing considerable confusion and implementation difficulties. Fortunately, the swarm of data standards applicable to IDI is beginning to narrow to a manageable group by the combined efforts of the National Center for Vital Health Statistics (NCVHS) and the Consolidated Health Informatics (CHI) E-Gov initiative (Goldsmith, Blumenthal et al. 2003). See Table 26-1 for some of the key standards.

Table 26-1. CHI Standards Applicable to IDI (Zeng, Chen et al. 2005)

CHI Adopted Standard	*Domain*
Health Level 7 (HL-7) messaging	messaging
Laboratory Logical Observation Identifier Name Codes (LOINC)	laboratory test orders
SNOMED CT	laboratory result contents; non-laboratory interventions and procedures, anatomy, diagnosis and problems
RxNORM	describing clinical drugs
HL-7 clinical vaccine formulation (CVX) and manufacturer codes (MVX)	immunization registry, terminology

While these standards are currently required only for federal government information systems, in all likelihood, data standards adopted by the federal government will be assimilated and adopted by private industry over a relatively short period of time due to the combination of payor (Medicare, Medicaid, and the Civilian Health and Medical Program of the Uniformed Services (CHAMPUS)) pressures, the sheer size of the federal government health care sector, and the need for private industry to communicate with these government systems. The Health Resources and Services Administration (HRSA) has also provided funds for encoding hospital laboratory information systems with the intention of helping migrate the systems from local code sets or CPT4 and ICD-9 code systems to LOINC and SNOMED

codes that will allow interoperability with local, state, and federal health information systems adhering to CHI standards.

In the public health sector, the CDC has led the way in the push for data standardization through the National Electronic Disease Surveillance System (NEDSS) and the Public Health Information Network (PHIN) initiatives. These initiatives define a set of vocabularies, messaging standards, message and data formats as well as the architectural components required for public health jurisdictions utilizing the federal Bioterrorism grants for funding information systems development. The National Library of Medicine brokered contract with the American College of Pathologists for the United States licensure of the SNOMED vocabulary, the naming of the first National Health Information Technology Coordinator, and the ongoing work on the National Health Information Infrastructure (NHII) provides the means for accelerating the pace for data standardization.

System Architecture and Messaging. There are fewer messaging standards relevant to IDI. Among them, Health Level 7 (HL7) is the dominant messaging standard for transfer of clinical information. Almost all hospital information systems exchange HL7 messages and the majority of large private clinical labs have adopted the HL7 standard as well. The current ANSI approved version of HL7 is 2.5; however, several new Version 3 messages for public health reporting have been developed and are being reviewed for implementation as a normative standard. The HL7 Version 3 specification represents a paradigm shift from the flat file structure of the 2.x HL7 versions to an object oriented Reference Information Model (RIM) foundation. This change provides the necessary structure to disambiguate the detailed information in the message and maintain the contextual relationships between data elements that are critical in infectious disease and bioterrorism system to system communication. The CDC has set a goal of using Version 3 messages for morbidity reporting from states to the CDC. Additionally, the HL7 Clinical Document Architecture (CDA) standard is being considered in a variety of reporting and data collection scenarios including the CDC Outbreak Management System.

Data Ingest and Access Control. Data ingest control is responsible for checking the integrity and authenticity of data feeds from the underlying information sources. Access control is responsible for granting and restricting user access to potentially sensitive data. Data ingest and access control is particularly important in IDI applications because of obvious data confidentiality concerns and data sharing requirements imposed by data contributors. Although ingest and access control issues are common in many application domains, IDI poses some unique considerations and requirements. In most other applications, a user is either granted or denied access to a particular information item. In IDI applications, however, user access

privilege is often not binary. For instance, a local public health official has full access to data collected from his or her jurisdiction but typically does not have the same access to data from neighboring jurisdictions. However, it does not necessarily mean that this official has no access at all to such data from neighboring jurisdictions. Often he or she can be granted access to such data in some aggregated form (e.g., monthly or county-level statistics). Such granularity-based data access requirements warrant special treatment when designing an IDI system.

3. SYNDROMIC SURVEILLANCE SYSTEMS

In the previous section, we briefly introduced the field of IDI and discussed the basic components of IDI systems. We noted that the discussion of many advanced functionalities of IDI systems needs to be framed in an application-specific manner. This section provides such an application context, i.e., syndromic surveillance, which allows us to have an extended discussion of these basic IDI system components along with advanced IDI data analysis and visualization techniques. Syndromic surveillance by itself represents a major trend in both research and real-world implementation and is arguably the most active and important IDI application in the current practice.

3.1 Background

Public health surveillance has been practiced for decades and continues to be an indispensable approach for detecting emerging disease outbreaks and epidemics. Early knowledge of a disease outbreak plays an important role in improving response effectiveness. While traditional disease surveillance often relies on time-consuming laboratory diagnosis and the reporting of notifiable diseases is often slow and incomplete, a new breed of public health surveillance systems has the potential to significantly speed up detection of disease outbreaks. These new, computer-based surveillance systems offer valuable and timely information to hospitals as well as to state, local, and federal health officials (Pavlin 2003; Bravata, McDonald et al. 2004; Dembek, Carley et al. 2005; Yan, Zeng et al. 2006). These systems are capable of real-time or near real-time detection of serious illnesses and potential bioterrorism agent exposures, allowing for a rapid public health response (Mandl, Overhage et al. 2004). This public health surveillance approach is generally called syndromic surveillance, an ongoing, systematic collection, analysis, and interpretation of “syndrome”-specific data for early detection of public health aberrations.

In the literature, the discussion of syndromic surveillance systems usually falls under the following functional areas: 1) data sources and acquisition, 2) syndrome classification, 3) anomaly detection, and 4) data visualization and data access. The surveillance data are first collected from the data providers to a centralized data repository where the raw data are categorized into syndrome categories to indicate certain disease threats. Anomaly detection employing time and space data analysis algorithms characterizes the syndromic data to detect the anomalies (for example, the surge of counts of clinic visits aggregated by days, or anomalous spatial clusters of medical records aggregated by ZIP codes). The data visualization and data access module is used to facilitate case investigations and support data exploration and summarization in a visual environment. The rest of this section is dedicated to these four functional areas, respectively.

3.2 Data Sources and Acquisition

Syndromic surveillance is a data-driven public health surveillance approach which collects and processes a wide array of data sources. These data sources include chief complaints, emergency department (ED) visits, ambulatory visits, hospital admissions, triage nurse calls, 911 calls, work or school absenteeism data, veterinary health records, laboratory test orders, and health department requests for influenza testing, among others (Lombardo, Burkom et al. 2004; Ma, Rolka et al. 2005). For instance, one of the most established syndromic surveillance projects, the Real-time Outbreak Detection system (RODS), uses laboratory orders, dictated radiology reports, dictated hospital reports, poison control center calls, chief complaint data, and daily sales data for over-the-counter (OTC) medications for syndromic surveillance (Tsui, Espino et al. 2003).

Preliminary investigations have evaluated the effectiveness of different data sources in syndromic surveillance and studied the difference among them in terms of information timeliness and characterization ability for outbreak detection, as they represent various aspects of patient health-care-seeking behavior (Lazarus, Kleinman et al. 2001; Ma, Rolka et al. 2005). For example, school absenteeism comes to notice relatively earlier as individuals take leave before seeking health care in hospitals or clinics, but specific disease evidence provided by the absenteeism type of data is limited. Table 26-2 provides a rough-cut classification of different data sources used for syndromic surveillance organized by their timeliness and the capability to characterize epidemic events.

Data acquisition is a critical early step when developing a syndromic surveillance system. The particular data collection strategy is obviously dependent on the data providers' information system infrastructure. Such

strategies range from direct manual entry on paper or using hand-held devices (Zelicoff, Brillman et al. 2001) to automated data transmission, archiving, query and messaging (Lombardo, Burkom et al. 2003; Espino, Wagner et al. 2004).

Table 26-2. Data sources organized by data timeliness and epidemic characterization (Yan, Zeng et al. 2006)

Timeliness	Characterization	
	High	Low
High	• Chief complaints from ED visits and ambulatory visits • Hospital admission • Test orders • Triage nurse calls, 911 calls • Prescription medication data	• OTC medication sales • School or work absenteeism • Veterinary health records
Low	• ICD-9 code • Laboratory test results • Clinical reports	• Public sources (local or regional events)

Many practical challenges exist that still hinder data collection efforts, including the following: (a) different coding conventions among the health facilities need to be reconciled when integrating the different data sources; (b) providing and transmission of data either requires staff intervention or dedicated network infrastructure with relatively high security level, which are often viewed as extra cost to data providers; (c) data sharing and transmission must comply with HIPAA and others, to be secure and assure privacy; and (d) there is a time lag getting data from data providers to syndromic surveillance systems. Data quality challenges (e.g., incompleteness and duplications) often pose additional challenges.

3.3 Syndrome Classification

The onset of a number of syndromes can indicate certain diseases that are viewed as threats to the public health. For example, influenza-like syndrome could be due to an anthrax attack, which is of particular interest in the detection of bioterrorism events.

A syndrome category is defined as a set of symptoms, which is an indicator of some specific diseases. For example, a short-phrase chief complaint "coughing with high fever" can be classified as the "upper respiratory" syndrome. Table 26-3 summarizes some of the most commonly-monitored syndrome categories. Note that different syndromic surveillance systems may monitor different categories. For example, in the RODS system there are 7 syndrome groups of interest for bio-surveillance purposes;

whereas EARS defines a more detailed list of 43 syndromes (http://www.bt.cdc.gov/surveillance/ears/). Some syndromes are of common interest across different systems, such as respiratory or gastrointestinal syndromes.

Table 26-3. Syndrome categories commonly monitored

Influenza-like	Respiratory	Dermatological
Fever	Neurologic	Cold
Gastrointestinal	Rash	Diarrhea
Hemorrhagic illness	Severe illness and death	Asthma
Localized cutaneous lesion	Specific infection	Vomit
Lymphadenitis	Sepsis	Other/none of the above
Constitutional		
Bioterrorism agent-related diseases		
Anthrax	Botulism-like/botulism	Plague
Tularemia	Smallpox	SARS (Severe Acute Respiratory Syndrome)

Currently the syndrome classification process is implemented into syndromic surveillance systems either manually or through an automated system. Note, however, automated, computerized syndrome classification is essential to real-time syndromic surveillance. Syndrome classification is thus one of the first and most important steps in syndromic data processing. The software application that evaluates the patient's chief complaint or ICD-9 codes and then assigns it to a syndrome category is often known as a syndrome classifier (Lu, Zeng et al. 2006).

A substantial amount of research effort has been expended to classifying free-text chief complaints into syndromes. This classification task is difficult because different expressions, acronyms, abbreviations, and truncations are often found in free-text chief complaints (Sniegoski 2004). For example, "chst pn," "CP," "c/p," "chest pai," "chert pain," "chest/abd pain," and "chest discomfort" can all mean "chest pain." As we observed in the previous section, a majority of syndromic surveillance systems use chief complaints as a major source of data; as a result, syndrome classification has wide applications. Another syndromic data type often used for syndromic surveillance purposes, i.e. ICD-9 or ICD-9-CM codes, also needs to be grouped into syndrome categories. Processing such information is somewhat easier as the data records are structured.

Classification methods that have been studied and employed can largely be categorized into four groups: 1) Natural language processing; 2) Bayesian classifiers; 3) Text string searching; and 4) Vocabulary abstraction. We summarize existing classification methods in Table 26-4.

Table 26-4. Syndrome classification approaches (Yan, Zeng et al. 2006)

Approach	Description	Example Systems
Manual grouping	Medical experts in syndromic surveillance, infectious diseases, and medical informatics perform the mapping of laboratory test orders into syndrome categories (Ma et al., 2005).	The BioSense system (Bradley et al., 2005; Sokolow et al., 2005) and Syndromal Surveillance Tally Sheet program used in EDs of Santa Clara County, California.
Natural language processing (NLP)	NLP-based approaches classify free-text CCs with simplified grammar containing rules for nouns, adjectives, prepositional phrases, and conjunctions. Critiques of NLP-based methods include lack of semantic markings in chief complaints and the amount of training needed.	As part of RODS, Chapman et al. adapted the MPLUS, a Bayesian network-based NLP system, to classify the free-text chief complaints (Chapman et al., 2005) (Chapman et al., 2005; Wagner, Espino et al., 2004).
Bayesian classifiers	Bayesian classifiers, including naïve Bayesian classifiers, bigram Bayes, and their variations, can classify CCs learned from the training data consisting of labeled CCs.	The CoCo Bayesian classifier from the RODS project (Chapman et al., 2003).
Text string searching	A rule-based method that first uses keyword matching and synonym lists to standardize CCs. Predefined rules are then used to classify CCs or ICD-9 codes into syndrome categories.	EARS (Hutwagner et al., 2003), ESSENCE (CDC, 2003), and the National Bioterrorism Syndromic Surveillance Demonstration Program (Yih et al., 2005).
Vocabulary abstraction	This approach creates a series of intermediate abstractions up to a syndrome category from the individual data (e.g., signs, lab tests) for syndromes indicative of illness due to an agent of bioterrorism.	The BioStorm system (Crubézy et al., 2005) (Buckeridge et al., 2002; Monica Crubézy, Martin O'Connor, Zachary Pincus, & Musen, 2005; Shahar & Musen, 1996).

*ESSENSE: Electronic Surveillance System for the Early Notification of Community-Based Epidemics; #EARS: Early Aberration Reporting System

Evaluations have been conducted to compare various classifiers' performance for certain selected syndrome types. For instance, experiments conducted on two Bayesian classifiers for acute gastrointestinal syndrome demonstrate a 68 percent mapping success against expert classification of ED reports (Ivanov, Wagner et al. 2002). Several technical challenges to syndromic classification remain. There are no standardized syndrome definitions employed universally by different syndromic surveillance systems. Different computerized classifiers, or human chief complaint coders, are trained to prioritize and code symptoms differently following different coding conventions. Studies demonstrate that the comparisons

between two syndrome coding systems show low agreement in most of the syndrome classifications (Mikosz, Silva et al. 2004).

3.4 Data Analysis and Outbreak Detection

Automated data analysis for aberration detection is essential to real-time syndromic surveillance. These algorithms, spanning from classic statistical methods to artificial intelligence approaches, are used to quantify the possibility of an outbreak observed from the surveillance data. For instance, such models have been employed to predict outbreaks of West Nile virus (Eidson, Miller et al. 2001; Wonham, de-Camino-Beck et al. 2004) and of influenza (Hyman and LaForce 2004).

Usually, a detection system employs more than one algorithm, as no single algorithm can cover the wide spectrum of possible situations. Below we sample representative algorithms organized as temporal, spatial, and spatial-temporal methods (Buckeridge, Burkom et al. 2005).

Another category, which is not shown in Table 26-5, includes "data-fusion" approaches where multiple data sources (e.g., ED visits and OTC sales data) are combined to perform outbreak detection. The idea of such "data-fusion" approaches is that multiple data streams are analyzed with the extensions of analytical techniques, such as MCUSUM or MEWMA, to increase detection sensitivity while limiting the number of false alarms.

Table 26-5. Outbreak Detection Algorithms

Algorithm	Short Description	Availability and Applications	Features and Problems
Temporal Analysis			
Serfling method	A static cyclic regression model with predefined parameters optimized through the training data	Available from RODS (Tsui, Wagner et al. 2001); used by CDC for flu detection; Costagliola *et al.* applied Serfling's method to the French influenza-like illness surveillance (Costagliola, Flahault et al. 1981)	The model fits data poorly during epidemic periods. To use this method the epidemic period has to be pre-defined.
Autoregressive Integrated Moving Average (ARIMA)	A linear function learns parameters from historical data. Seasonal effect can be adjusted.	Available from RODS	Suitable for stationary environments.

(Continued)

Table 26-5. (Continued)

Algorithm	Short Description	Availability and Applications	Features and Problems
Exponentially Weighted Moving Average (EWMA)	Predictions based on exponential smoothing of previous several weeks of data with recent days having the highest weight (Neubauer 1997)	Available from ESSENCE	Allowing the adjustment of shift sensitivity by applying different weighting factors.
Cumulative Sums (CUSUM)	A control chart-based method to monitor for the departure of the mean of the observations from the estimated mean (Das, Weiss et al. 2003; Grigoryan, Wagner et al. 2005). It allows for limited baseline data.	Widely used in current surveillance systems including BioSense, EARS (Hutwagner, Thompson et al. 2003) (Hutwagner, Thompson et al. 2003) and ESSENCE, among others	This method performs well for quick detection of subtle changes in the mean (Roger-son 2005); it is criticized for its lack of adjustability for seasonal or day-of-week effects.
Hidden Markov Models (HMM)	HMM-based methods use a hidden state to capture the presence or absence of an epidemic of a particular disease and learn probabilistic models of observations conditioned on the epidemic status.	Discussed in (Rath, Carreras et al. 2003)	A flexible model that can adapt automatically to trends, seasonality covariates (e.g., gender and age), and different distributions (normal, Poisson, etc.).
Wavelet algorithms	Local frequency-based data analysis methods; they can automatically adjust to weekly, monthly, and seasonal data fluctuations.	Used in NRDM to indicate zip-code areas in which OTC medication sales are substantially increased (Espino and Wagner 2001; Zhang, Tsui et al. 2003)	Account for both long-term (e.g., seasonal effects) and short-term trends (e.g., day-of-week effects) (Wagner, Tsui et al. 2004).
Spatial Analysis			
Generalized Linear Mixed Modeling (GLMM)	Evaluating whether observed counts in relatively small areas are larger than expected on the basis of the history of naturally occurring diseases (Kleinman, Lazarus et al. 2004; Kleinman, Abrams et al. 2005)	Used in Minnesota (Yih, Abrams et al. 2005)	Sensitive to a small number of spatially focused cases; poor in detecting elevated counts over contiguous areas as compared to scan statistic and spatial CUSUM approaches (Kleinman, et al. 2004).

(Continued)

Table 26-5. (Continued)

Algorithm	Short Description	Availability and Applications	Features and Problems
Spatial scan statistics and variations	The basic model relies on using simply-shaped areas to scan the entire region of interest based on well-defined likelihood ratios. Its variation takes into account factors such as people mobility	Widely adopted by many syndromic surveillance systems; a variation proposed in (Duczmal and Buckeridge 2005); visualization available from BioPortal (Zeng, Chang et al. 2004).	Well-tested for various outbreak scenarios with positive results; the geometric shape of the hotspots identified is limited.
Bayesian spatial scan statistics	Combining Bayesian modeling techniques with the spatial scan statistics method; outputting the posterior probability that an outbreak has occurred, and the distribution of this probability over possible outbreak regions	Available from RODS (Neill, Moore et al. 2005)	Computationally efficient; can easily incorporate prior knowledge such as the size and shape of outbreak or the impact on the disease infection rate.
Spatial-temporal Analysis			
Space-time scan statistic	An extension of the space scan statistic that searches all the sub-regions for likely clusters in space and time with multiple likelihood ratio testing (Kulldorff 2001).	Widely used in many community surveillance systems including the National Bioterrorism Syndromic Surveillance Demonstration Program (Yih, Caldwell et al. 2004)	Regions identified may be too large in coverage.
What is Strange About Recent Event (WSARE)	Searching for groups with specific characteristics (e.g., a recent pattern of place, age, and diagnosis associated with illness that is anomalous when compared with historic patterns) (Kaufman, Cohen et al. 2005)	Available from RODS; Implemented in ESSENCE	In contrast to traditional approaches, this method allows for use of representative features for monitoring (Wong, Moore et al. 2002; Wong, Moore et al. 2003). To use it, however, the base-line distribution has to be known.

(Continued)

Table 26-5. (Continued)

Algorithm	Short Description	Availability and Applications	Features and Problems
Prospective Support Vector Clustering (PSVC)	This method uses the Support Vector Clustering method with risk adjustment as a hotspot clustering engine and a CUSUM-type design to keep track of incremental changes in spatial distribution patterns over time	Developed in BioPortal (Zeng, Chang et al. 2004; Chang, Zeng et al. 2005)	This method can identify hotspots with irregular shapes in an online context

Another challenging issue for real time outbreak detection is that the surveillance algorithms rely on historic datasets that span a considerable length of time against which to measure the anomaly of observed data. Few methods demonstrate reliable detection capability with short-term baseline data. Measurements on timeliness, specificity, and sensitivity of the detection algorithms have been reported; however, existing evaluation studies are quite limited, as they are either reliable only for one specific disease (Kleinman and Abrams 2006) or are biased by simulation settings as very few bioterrorism attacks for real testing are available.

3.5 Data Access and Visualization

To facilitate interactive data exploration, maps, graphs, and tables are common forms of helpful visualization tools. Below we briefly review some sample implementations in IDI contexts. Section 4 contains several more detailed case studies with screenshots.

The RODS system employs the GIS module to depict data spatially. In BioSense and ESSENSE, a geographical map consisting of individual zip codes is marked with different colors to represent the threat level. Stratification can be applied for different syndrome categories, and individual case details can be accessed by "drill down" functions. The BioPortal project makes available an advanced visualization module, called the Spatial Temporal Visualizer (STV) to facilitate exploration of infectious disease case data and to summarize query results (Hu, Zeng et al. 2005). STV is a generic visualization environment that can be used to visualize a number of spatial temporal datasets simultaneously. It allows the user to load and save spatial temporal data in a dynamic manner for exploration and dissemination.

4. IDI AND SYNDROMIC SURVEILLANCE SYSTEM CASE STUDIES

To better illustrate the earlier discussion on the data sources used and technical components of IDI systems, and related implementation issues, we present two case studies in this section.

4.1 RODS

The first case study examines the Realtime Outbreak and Disease Surveillance (RODS) system, which has been deployed across the nation. The RODS project is a collaborative effort between the University of Pittsburgh and Carnegie Mellon University. It provides a computing platform for the implementation and evaluation of different analytic approaches for outbreak detection, among other data collection and reporting functions.

The RODS project was initiated by the RODS Laboratory in 1999. The system is now an open source project under the GNU license. The RODS development effort has been organized into seven functional areas: overall design, data collection, syndrome classification, database and data warehousing, outbreak detection algorithms, data access, and user interfaces. Each functional area has a coordinator for the related open source project effort and there is an overall coordinator responsible for the architecture, overall integration of components, and overall quality of the source code. Figure 26-1 illustrates the RODS' system architecture.

The RODS system as a syndromic surveillance application was originally deployed in Pennsylvania, Utah, and Ohio. It is currently deployed in New Jersey, Michigan, and several other states. By June 2006, about 20 regions with more than 200 health care facilities connected to RODS in real-time. It was also deployed during the 2002 Winter Olympics (Espino, Wagner et al. 2004).

The RODS data are collected in real-time through HL7 messages from other computer systems such as registration systems and laboratory information systems, over a Secure Shell–protected Internet connection in an automated mode. The National Retail Data Monitor (NRDM) is a component of the RODS system, collecting and analyzing daily sales data for OTC medication sales. It also collects and analyzes chief complaint data from various hospitals. The RODS system currently monitors 8 syndrome categories: Gastrointestinal, Hemorrhagic illness, Constitutional, Neurologic, Rash, Respiratory, Botulism-like/botulism, Others.

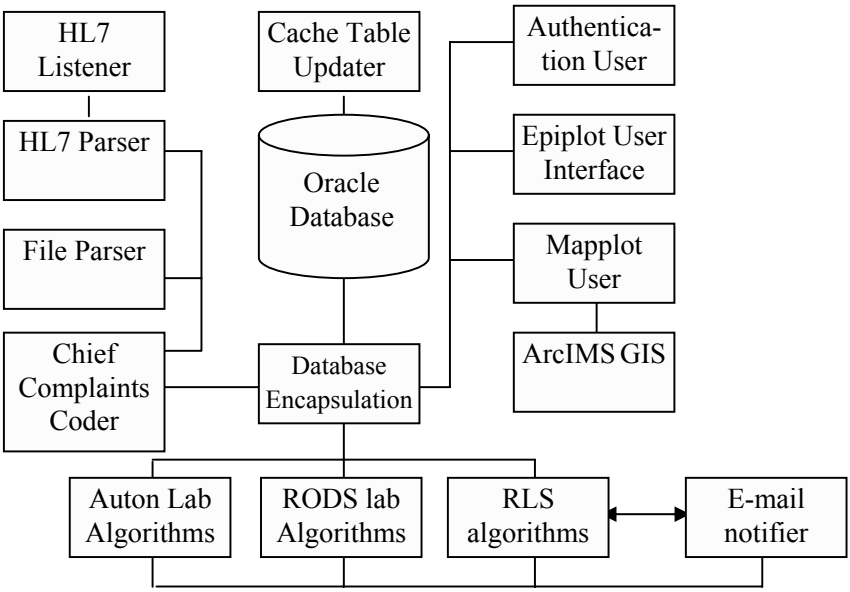

Figure 26-1. RODS system architecture (Espina, Wagner et al. 2004)

One of the major strengths of RODS is in data analysis. A number of syndrome classification approaches have been tested and implemented in the RODS system. It applies a keyword classifier and an ICD-9 classifier to chief complaint data. The CoCo module, a syndrome mapping component, has been tested in multiple settings (Olszewski 2003). For the respiratory syndrome, based on manually-classified results, CoCo's sensitivity level achieves 77% and specificity level 90%. A Bayesian network-based semantic model has been shown to classify free-text chief complaints effectively at the expense of added system complexity and computational overhead (Chapman, Christensen et al. 2005). The performance of the classifier represented by the ROC curve for each syndrome category varies between 0.95 and 0.99.

The current open source release of the RODS system includes implementations of several outbreak detection algorithms: wavelet-detection algorithms, CUSUM, SMART, scan statistics, RLS, and WSARE.

RODS provides multiple graphing techniques with both time-series and geographical displays available via an encrypted, password-protected Web interface. Three different data views — Main, Epiplot, and Mapplot — are supported. The main RODS screen shows time-series plots updated on a daily basis for each syndrome. The user can also view these graphs by county or for the whole state. The Epiplot screen is highly interactive; the user can specify the syndrome, region, start dates, and end dates, to generate

customized time-series plots. A "get cases" button allows users to view case-level detail for encounters making up the specific time-series. The Mapplot screen provides an interface to the ArcIMS package, to display disease cases' spatial distribution using patients' zip code information.

4.2 BioPortal

The second case study is about the BioPortal system. The BioPortal project was initiated in 2003 by the University of Arizona and its collaborators in the New York State Department of Health and the California Department of Health Services under the guidance of a federal inter-agency working group named the Infectious Disease Informatics Working Committee. Since then, its partner base has expanded to include the USGS, University of California, Davis, University of Utah, the Arizona Department of Health Services, Kansas State University, and the National Taiwan University. BioPortal provides distributed, cross-jurisdictional access to datasets concerning several major infectious diseases, including Botulism, West Nile virus, foot-and-mouth disease, livestock syndromes, and chief complaints (both in English and Chinese). It features advanced spatial-temporal data analysis methods and visualization capabilities.

Figure 26-2 shows its system architecture. BioPortal supports syndromic surveillance of epidemiological data and free-text chief complaints. It also supports analysis and visualization of lab-generated gene sequence information.

Figure 26-3 illustrates how epidemiological and genetic data analyses are integrated from a systems perspective.

As to data collection, emergency room chief complaint data in the free-text format are provided by the Arizona Department of Health Services and several hospitals in batch mode for syndrome classification. Various disease-specific case reports for both human and animal diseases are another source of data for BioPortal. It also makes use of surveillance datasets such as dead bird sightings and mosquito control information. The system's communication backbones, initially for data acquisition from New York or California disease datasets, consist of several messaging adaptors that can be customized to interoperate with various messaging systems. Participating syndromic data providers can link to the BioPortal data repository via the PHINMS and an XML/HL7 compatible network.

BioPortal provides automatic syndrome classification capabilities based on free-text chief complaints. One method recently developed uses a concept ontology derived from the UMLS (Lu, Zeng et al. 2006). For each chief complaint (CC), the method first standardizes the CC into one or more medical concepts in the UMLS. These concepts are then mapped into exist-

ing symptom groups using a set of rules constructed from a symptom grouping table. For symptoms not in the table, a Weighted Semantic Similarity Score algorithm, which measures the semantic similarity between the target symptoms and existing symptom groups, is used to determine the best symptom group for the target symptom. The ontology-enhanced CC classification method has also been extended to handle CCs in Chinese.

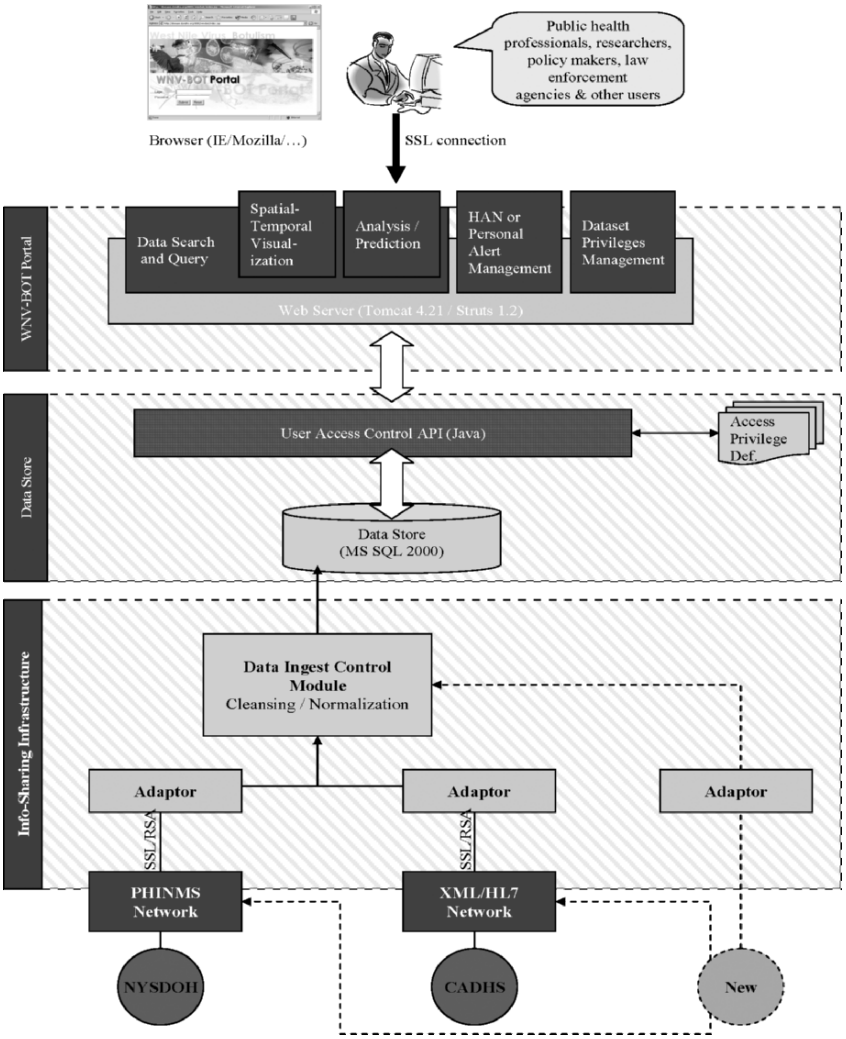

Figure 26-2. BioPortal information sharing and data access infrastructure

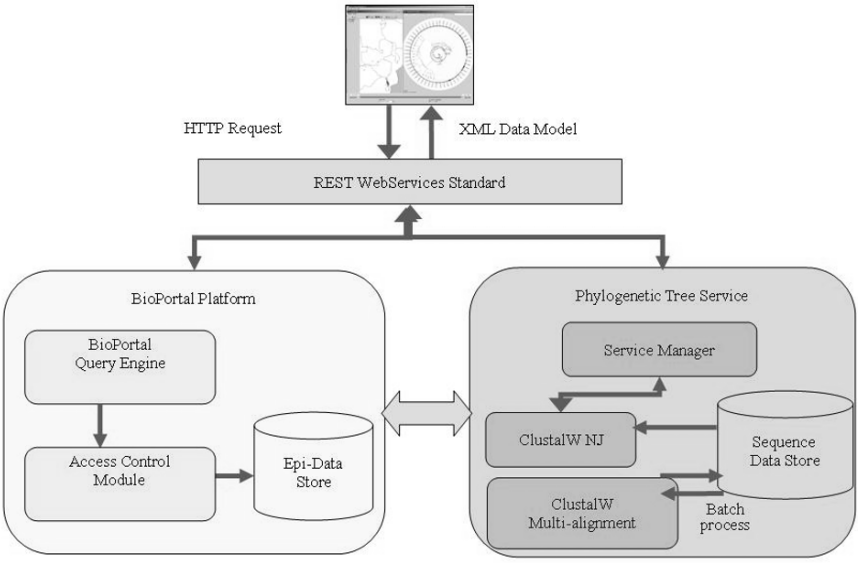

Figure 26-3. Integrating epidemiological data and gene sequence data analysis in BioPortal

BioPortal supports outbreak detection based on spatial-temporal clustering analysis, also known as hotspot analysis, to identify unusual spatial and temporal clusters of disease events. BioPortal supports various scan statistic-based methods through the SaTScan binary, which has been widely used in public health; the Nearest Neighbor Hierarchical Clustering method initially developed for crime analysis; and two new artificial intelligence clustering-based methods (Risk-Adjusted Support Vector Clustering, and Prospective Support Vector Clustering) developed in-house, which can support detection of areas with irregular shapes (Zeng, Chang et al. 2004; Chang, Zeng et al. 2005).

BioPortal makes available a visualization environment called the Spatial-Temporal Visualizer (STV), which allows users to interactively explore spatial and temporal patterns, based on an integrated tool set consisting of a GIS view, a timeline tool, and a periodic pattern tool.

Figure 26-4 illustrates how these three views can be used to explore an infectious disease dataset. The GIS view displays cases and sightings on a map. The user can select multiple datasets to be shown on the map in different layers using the checkboxes (e.g., disease cases, natural land features, and land-use elements). Through the periodic view the user can identify periodic temporal patterns (e.g., which months or weeks have an unusually high number of cases). The unit of time for aggregation can also be set as days or hours. The timeline view provides a timeline along with a

hierarchical display of the data elements, organized as a tree. A new gene sequence-based phylogenetic tree visualizer (not shown in Figure 26-4) has been recently added to the STV interface for diseases with quick mutation rates such as the foot-and-mouth disease. This allows BioPortal users to explore epidemiological and sequence data concurrently.

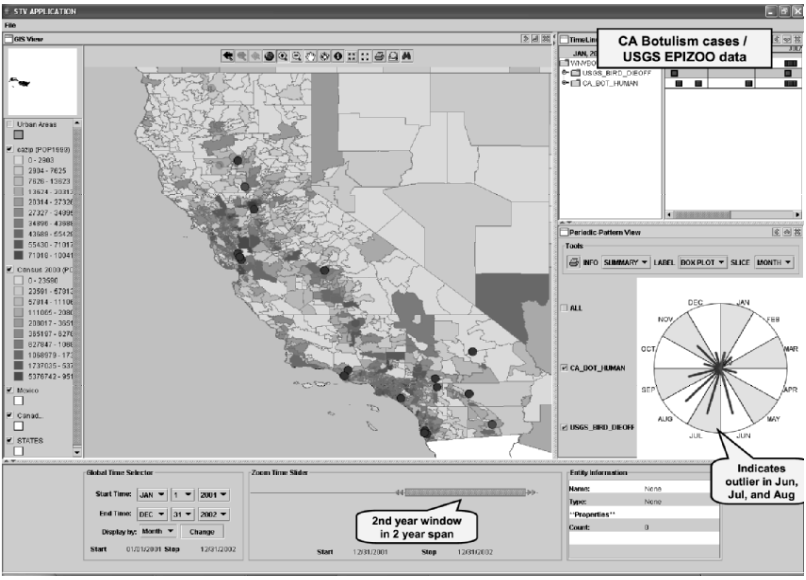

Figure 26-4. Spatial-temporal visualizer

Data confidentiality, security, and access control are among the key research and development issues for the BioPortal project. A role-based access control mechanism is implemented based on data confidentiality and user access privileges. The project has also developed a consortium type of data sharing Memoranda of Understanding (MOU) to reduce the barrier of sharing information among data contributors including local and state public health agencies.

5. CONCLUDING REMARKS

This chapter provides a brief review of infectious disease informatics (IDI) with a particular focus on its application in syndromic surveillance. Traditional disease surveillance systems are based on confirmed diagnoses, whereas syndromic surveillance makes use of pre-diagnosis information for

timely data collection and analysis. The main IT-related topics and challenges in IDI and syndromic surveillance are presented in this chapter.

With regards to ongoing trends in IDI and syndromic surveillance, we see significant interest in informatics studies on topics ranging from data visualization, further development and comprehensive evaluation of outbreak detection algorithms, data interoperability, and further development of response and event management decision models based on data and predictions provided by syndromic surveillance systems.

From an application domain perspective, the following areas can potentially lead to new and interesting innovations and research. First, public health surveillance can be a truly global effort for pandemic diseases such as avian influenza. Issues concerning global data sharing (including multilingual information processing) and the development of models that work over a wide geographical area need to be addressed. Syndromic surveillance concepts, techniques, and systems are equally applicable to animal health besides public health. We expect to see significant new work to be done to model animal health-specific issues and deal with zoonotic diseases.

ACKNOWLEDGEMENTS

This work is supported in part by the U.S. National Science Foundation through Digital Government Grant #EIA-9983304, Information Technology Research Grant #IIS-0428241, by the U.S. Department of Agriculture through Grant #2006-39546-17579, and by the Arizona Department of Health Services. We would like to thank the members of the BioPortal project for insightful discussions. The first author is an affiliated professor at the Institute of Automation, the Chinese Academy of Sciences, and wishes to acknowledge support from a research grant (60573078) from the National Natural Science Foundation of China, an international collaboration grant (2F05N01) from the Chinese Academy of Sciences, and a National Basic Research Program of China (973) grant (2006CB705500) from the Ministry of Science and Technology.

REFERENCES

Bradley, C. A., H. Rolka, et al. (2005). "BioSense: Implementation of a National Early Event Detection and Situational Awareness System." MMWR (CDC) 54(Suppl): 11-20.

Bravata, D., K. McDonald, et al. (2004). "Systematic review: Surveillance Systems for Early Detection of Bioterrorism-Related Diseases." Annals of Internal Medicine 140: 910-922.

Buckeridge, D., H. Burkom, et al. (2005). "Algorithms for Rapid Outbreak Detection: a Research Synthesis." Journal of Biomedical Informatics 38: 99-113.

Buckeridge, D., J. Graham, et al. (2002). Knowledge-Based Bioterrorism Surveillance. American Medical Informatics Association Symposium, San Antonio, TX.

Buehler, J., R. Hopkins, et al. (2004). "Framework for evaluating public health surveillance systems for early detection of outbreaks: Recommendations from the CDC working group." MMWR (CDC) 53(RR-5): 1-13.

Chang, W., D. Zeng, et al. (2005). A Novel Spatio-Temporal Data Analysis Approach based on Prospective Support Vector Clustering. Proceedings of the Fifteenth Annual Workshop on Information Technologies and Systems (WITS'05), Las Vegas, NV.

Chang, W., D. Zeng, et al. (2005) "Prospective Spatio-Temporal Data Analysis for Security Informatics." Proceedings of the 8th IEEE International Conference on Intelligent Transportation Systems Vienna, Austria.

Chapman, W. W., L. Christensen, et al. (2005). "Classifying Free-text Triage Chief Complaints into Syndromic Categories with Natural Language Processing." Artificial Intelligence in Medicine 33(1): 31-40.

Christensen, L., P. Haug, et al. (2002). MPLUS: a Probabilistic Medical Language Understanding System. Workshop on Natural Language Processing in the Biomedical Domain.

Cooper, G. F., D. H. Dash, et al. (2004). Bayesian Biosurveillance of Disease Outbreaks. Twentieth Conference on Uncertainty in Artificial Intelligence, Banff, Alberta, Canada.

Costagliola, D., A. Flahault, et al. (1981). "A routine tool for detection and assessment of epidemics of influenza-like syndromes in France" American Journal of Public Health 81(1): 97-99.

Crubézy, M., M. O'Connor, et al. (2005). "Ontology-Centered Syndromic Surveillance for Bioterrorism." IEEE INTELLIGENT SYSTEMS 20(5): 26-35.

Das, D., D. Weiss, et al. (2003). "Enhanced Drop-in Syndromic Surveillance in New York City Following September 11, 2001." J Urban Health 80(1(suppl)): 176-188.

Dembek, Z., K. Carley, et al. (2005). "Guidelines for Constructing a Statewide Hospital Syndromic Surveillance Network." MMWR (CDC) 54(Suppl): 21-26.

Duczmal, L. and D. Buckeridge (2005). "Using Modified Spatial Scan Statistic to Improve Detection of Disease Outbreak When Exposure Occurs in Workplace — Virginia, 2004." MMWR (CDC) 54(Suppl): 187.

Eidson, M., J. Miller, et al. (2001). "Dead Crow Densities and Human Cases of West Nile Virus, New York State, 2000." Emerging Infectious Diseases 7: 662-664.

Espino, J. U. and M. M. Wagner (2001). "The Accuracy of ICD-9 Coded Chief Complaints for Detection of Acute Respiratory Illness." Proc AMIA Symp: 164-168.

Espino, J. U., M. M. Wagner, et al. (2004). "Removing a Barrier to Computer-Based Outbreak and Disease Surveillance — The RODS Open Source Project." MMWR (CDC) 53(Suppl): 34-41.

Goldsmith, J., D. Blumenthal, et al. (2003). "Federal Health Insurance Policy: A Case of Arrested Development." Health Affairs 22(4): 44-55.

Grigoryan, V. V., M. M. Wagner, et al. (2005) "The Effect of Spatial Granularity of Data on Reference Dates for Influenza Outbreaks." RODS Laboratory Technical Report, 2005

Hu, P.-J. H., D. Zeng, et al. (2007). "A System for Infectious Disease Information Sharing and Analysis: Design, Implementation and Evaluation." IEEE Transactions on Information Technology in Biomedicine in press.

Hu, P. J.-H., D. Zeng, et al. (2005). Evaluating an Infectious Disease Information Sharing and Analysis System. Intelligence and Security Informatics (ISI), Atlanta, Georgia, USA.

Hutwagner, L., W. Thompson, et al. (2003). "The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS)." J Urban Health 80(2 suppl 1): 89-96.

Hyman, J. and T. LaForce (2004). Modeling the spread of influenza among cities. Bioterrorism: Mathematical Modeling Applications in Homeland Security. H. Banks and C. Castillo-Chàvez, Society for Industrial and Applied Mathematics: 211-236.

Ivanov, O., M. M. Wagner, et al. (2002). Accuracy of Three Classifiers of Acute Gastrointestinal Syndrome for Syndromic Surveillance. AMIA Symp.

Kargupta, H., K. Liu, et al. (2003). Privacy Sensitive Distributed Data Mining from Multi-Party Data. Proc. of the first NSF/NIJ Symposium on Intelligence and Security Informatics, Springer LNCS 2665.

Kaufman, Z., E. Cohen, et al. (2005). "Using Data on an Influenza B Outbreak To Evaluate a Syndromic Surveillance System — Israel, June 2004 [abstract]." MMWR (CDC) 54(Suppl): 191.

Kleinman, K. and A. Abrams (2006). "Assessing surveillance using sensitivity, specificity and timeliness." Statistical Methods in Medical Research, Vol. 15, No. 5, 445-464 (15(5): 445-464.

Kleinman, K., A. Abrams, et al. (2005). "A Model-adjusted Spacetime Scan Statistic with an Application to Syndromic Surveillance." Epidemiol Infect 2005(119): 409-19.

Kleinman, K., R. Lazarus, et al. (2004). "A Generalized Linear Mixed Models Approach for Detecting Incident Cluster/signals of Disease in Small Areas, with an Application to Biological Terrorism (with Invited Commentary)." Am J Epidemiol 2004 159: 217-24.

Kulldorff, M. (2001). "Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic." Journal of the Royal Statistical Society, Series A(164): 61-72.

Lawson, A. B. and K. Kleinman, Eds. (2005). Spatial and Syndromic Surveillance for Public Health, John Wiley & Sons.

Lazarus, R., K. Kleinman, et al. (2001). "Using Automated Medical Records for Rapid Identification of Illness Syndromes (Syndromic Surveillance): the Example of Lower Respiratory Infection." BMC Public Health 1(9).

Lombardo, J., H. Burkom, et al. (2003). "A systems overview of the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE II)." Journal of Urban Health: Bulletin of the New York Academy of Medicine 80(2): 32-42.

Lombardo, J., H. Burkom, et al. (2004). "Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II), Framework for Evaluating Syndromic Surveillance Systems." Syndromic surveillance: report from a national conference, 2003. MMWR 2004 53(Suppl): 159-165.

Lu, H.-M., D. Zeng, et al. (2006). Ontology-based Automatic Chief Complaints Classification for Syndromic Surveillance. Proceedings of the IEEE Systems, Man, and Cybernetics, Taipei, IEEE Press.

Ma, H., H. Rolka, et al. (2005). "Implementation of Laboratory Order Data in BioSense Early Event Detection and Situation Awareness System." MMWR (CDC) 54(Suppl): 27-30.

Mandl, K. D., J. M. Overhage, et al. (2004). "Implementing Syndromic Surveillance: a Practical Guide Informed by the Early Experience." J Am Med Inform Assoc. 11(2): 141-50.

Mikosz, C. A., J. Silva, et al. (2004). "Comparison of Two Major Emergency Department-Based Free-Text Chief-Complaint Coding Systems." MMWR (CDC) 53(Suppl).

Moore, A. W., G. Cooper, et al. (2002) "Summary of Biosurveillance-relevant Statistical and Data Mining Techniques." RODS Laboratory Technical Report Volume, DOI:

Neill, D., A. Moore, et al. (2005). "A Bayesian Spatial Scan Statistic." Accepted to Neural Information Processing Systems 18.

Neubauer, A. (1997). "The EWMA Control Chart: Properties and Comparison with Other Quality-control Procedures by Computer Simulation." Clinical Chemistry 43(4): 594-601.

Ohno-Machado, L., P. S. S. PS, et al. (2004). "Protecting Patient Privacy by Quantifiable Control of Disclosures in Disseminated Databases." International Journal of Medical Informatics 73(7-8): 599-606.

Olszewski, R. T. (2003). Bayesian Classification of Triage Diagnoses for the Early Detection of Epidemics. 16th Int FLAIRS Conference.

Pavlin, J. A. (2003). "Investigation of Disease Outbreaks Detected by "Syndromic" Surveillance Systems." Journal of Urban Health: Bulletin of the New York Academy of Medicine 80(2): 107-114.

Pinner, R., C. Rebmann, et al. (2003). "Disease Surveillance and the Academic, Clinical, and Public Health Communities." Emerging Infectious Diseases 9(7): 781-787.

Rath, T. M., M. Carreras, et al. (2003). Automated Detection of Influenza Epidemics with Hidden Markov Models. Lecture Notes in Computer Science Springer Berlin/ Heidelberg 521 - 532

Rogerson, P. A. (2005). Spatial Surveillance and Cumulative Sum Methods. Spatial & Syndromic Surveillance for Public Health. K. K. Andrew B Lawson, John Wiley & Sons, Ltd: 95-113.

Siegrist, D. (1999). "The Threat of Biological Attack: Why Concern Now?" Emerging Infectious Diseases 5(4): 505-508.

Sniegoski, C. A. (2004). "Automated Syndromic Classifi cation of Chief Complaint Records." JOHNS HOPKINS APL TECHNICAL DIGEST 25(1): 68-75.

Sokolow, L. Z., N. Grady, et al. (2005). "Deciphering Data Anomalies in BioSense." MMWR (CDC) 54(Suppl): 133-140.

Tsui, F.-C., J. U. Espino, et al. (2003). "Technical Description of RODS: a Real-time Public Health Surveillance System." J Am Med Inform Assoc 2003 10: 399-408.

Tsui, F.-C., M. M. Wagner, et al. (2001). "Value of ICD-9–Coded Chief Complaints for Detection of Epidemics." Symposium of Journal of American Medical Informatics Association.

Wagner, M. M., A. W. Moore, et al. (2006). Handbook of Biosurveillanec, Elsevier.

Wagner, M. M., F.-C. Tsui, et al. (2004). "National Retail Data Monitor for Public Health Surveillance." MMWR (CDC) 53(Suppl): 40-42.

Wong, W. K., A. Moore, et al. (2003). "WSARE: What's Strange about Recent Events?" Journal of Urban Health 80((2 Suppl. 1)): 66-75.

Wong, W. K., A. Moore, et al. (2002). Rule-based Anomaly Pattern Detection for Detecting Disease Outbreaks. AAAI-02, Edmonton, Alberta

Wonham, M., T. de-Camino-Beck, et al. (2004). "An epidemiological model for West Nile virus: invasion analysis and control applications." Proceedings of Royal Society: Biological Sciences 271(1538): 501-507.

Wylie, J. E. and G. P. Mineau (2003). "Biomedical Databases: Protecting Privacy and Promoting Research." Trends in Biotechnology 21(3): 113-116.

Yan, P., D. Zeng, et al. (2006). A Review of Public Health Syndromic Surveillance Systems. ISI 2006, San Diego, CA, USA, Springer.

Yih, W., B. Caldwell, et al. (2004). "The National Bioterrorism Syndromic Surveillance Demonstration Program." MMWR (CDC) 53(Suppl): 43-6.

Yih, W. K., A. Abrams, et al. (2005). "Ambulatory-Care Diagnoses as Potential Indicators of Outbreaks of Gastrointestinal Illness — Minnesota." MMWR (CDC) 54(Suppl): 157-162.

Zelicoff, A., J. Brillman, et al. (2001). The rapid syndrome validation project (RSVP). AMIA Symp.

Zeng, D., W. Chang, et al. (2004). Clustering-based Spatio-Temporal Hotspot Analysis Techniques in Security Informatics. IEEE Transactions on Intelligent Transportation Systems.

Zeng, D., W. Chang, et al. (2004). A Comparative Study of Spatio-Temporal Data Analysis Techniques in Security Informatics. Proceedings of the 7th IEEE International Conference on Intelligent Transportation Systems, Washington, DC.

Zeng, D., H. Chen, et al. (2005). Disease Informatics and Outbreak Detection. Advances in Medical Informatics: Knowledge Management and Data Mining in Biomedicine. H. Chen, S. Fuller and A. McCray, Springer.

Zeng, D., H. Chen, et al. (2004). Towards A National Infectious Disease Information Infrastructure: A Case Study in West Nile Virus and Botulism. Proceedings of National Conference on Digital Government.

Zeng, D., H. Chen, et al. (2004). West Nile virus and botulism portal: A case study in infectious disease informatics. Intelligence and Security Informatics, ISI-2004, Lecture Notes in Computer Science.

Zhang, J., F. Tsui, et al. (2003). "Detection of Outbreaks from Time Series Data Using Wavelet Transform." AMIA Symp 748-52.

SUGGESTED READINGS

- "Handbook of Biosurveillance," M. M. Wagner, A. W. Moore, R. M. Aryel (eds.), 2006. Academic Press. This edited volume provides a comprehensive review of the theory and practice of real-time human disease outbreak detection. Its discussion on statistical outbreak detection methods is particularly informative.
- "Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance," R. Brookmeye and D. F. Stroup (eds.), 2004. Oxford University Press. This text is primarily concerned with statistical models as applied to health surveillance data and related analysis and interpretations challenges.
- "Public Health Informatics and Information Systems." P. O'Carroll, W. Yasnoff, E. Ward, L. Ripp, and E. Martin (eds.), 2002. Springer. This

edited book covers all aspects of public health informatics and presents a strategic approach to information systems development and management.

- "Bioterrorism: Mathematical Modeling Applications in Homeland Security," H.T. Banks and C. Castillo-Chàvez (eds.), 2003. The Society for Industrial and Applied Mathematics. This edited volume covers recent research on bio-surveillance, agroterrorism, bioterrorism response logistics, and assessment of the impact of bioterrorism attacks. The specific emphasis of this book is on mathematical modeling and computational studies relevant to bioterrorism research.

ONLINE RESOURCES

- CDC's NEDSS homepage at http://www.cdc.gov/nedss/index.htm
- CDC's PHIN homepage at http://www.cdc.gov/phin/, including the BioSense project homepage at http://www.cdc.gov/PHIN/component-initiatives/biosense/index.html
- Health Level Seven standards and software implementation at http://www.hl7.org
- Scan statistics-related outbreak detection software, datasets, and selected publications http://www.satscan.org
- The RODS project's homepage at http://rods.health.pitt.edu/
- The ESSENCE project's homepage at: http://www.geis.fhp.osd.mil/aboutGEIS.asp.
- The BioPortal project's homepage at http://www.bioportal.org
- http://statpages.org/ and http://www.autonlab.org/tutorials/, for related statistical and data mining tutorials.

QUESTIONS FOR DISCUSSION

1. What patient confidentiality, and data ownership and access control issues need to be considered in the IDI context?
2. What is the current status of IDI data and messaging standard development? What role should government play in the standardization effort?

3. Scan statistics and hotspot analysis techniques can identify unusual clustering of events or cases in space and time. How can one interpret the findings based on these techniques in the IDI context?
4. What role can visualization play in IDI data analysis? What are the types of visualizations commonly used by public health officials (not necessarily computerized) in disease surveillance?
5. What are the potential policy and organizational barriers to the deployment of syndromic surveillance systems? How can we overcome these barriers?
6. What are the technical obstacles associated with developing a syndromic surveillance system with international coverage? What are the non-technical obstacles?