Chapter 3

SYNDROMIC SURVEILLANCE DATA SOURCES AND COLLECTION STRATEGIES

In this and the ensuing two chapters, we will focus on three key technical aspects of modern syndromic surveillance systems: data sources and collection strategies; data analysis and outbreak detection; and data visualization, information dissemination, and reporting.

This chapter discusses syndromic data collection strategies and related data sources. Data collection is a critical early step when developing a syndromic surveillance system. It involves the selection of data sources, choices over vocabulary to be used, data entry approaches, and data transmission strategies and protocols. We will go through the related technical issues in the following sections. Towards the end of this chapter, we briefly summarize additional policy-related considerations that may impact data collection.

1. DATA SOURCES FOR PUBLIC HEALTH SYNDROMIC SURVEILLANCE

Syndromic surveillance is a largely data-driven public health surveillance approach. Data sources used in syndromic surveillance systems are expected to provide timely, prediagnosis health indicators and are typically electronically stored and transmitted. Note that most syndromic surveillance data were originally collected and used for other purposes and such data now serve dual purposes. Figure 3-1 depicts the conceptual timeline of prediagnosis data types and sources for syndromic surveillance.

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_3,

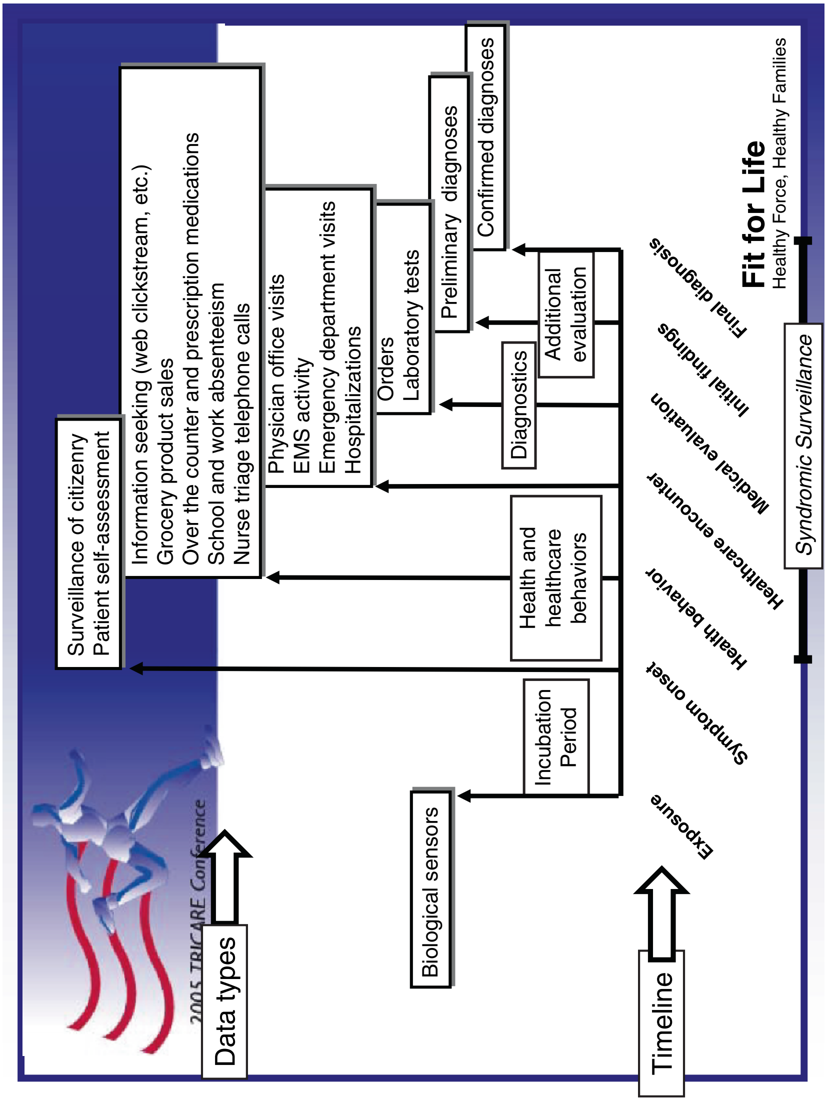

Figure 3-1. Conceptual timeline of collection and analysis of prediagnosis information for Syndromic Surveillance (2005c).

According to an empirical study conducted by Platt et al. (2003), most data collected for syndromic surveillance purposes include similar data elements: demographic data such as gender, age, area of residence; and data relevant to patient visits such as hospital name, the date of the visit, and the symptom set (chief complaints or admission status).

In this monograph, we identify the range of syndromic data sources and briefly summarize how they are used. Healthcare providers, schools, pharmacies, laboratories, and military medical facilities are all data contributors for syndromic surveillance. Specifically, data used for syndromic surveillance include emergency department (ED) visit chief complaints, ambulatory visit records, hospital admissions, OTC drug sales from pharmacy stores, triage nurse calls, 911 calls, work or school absenteeism data, veterinary health records, laboratory test orders, and health department requests for influenza testing (Ma et al., 2005).

Chief complaints record patient-reported signs and symptoms of their illness (e.g., coughing, headache, etc.) for ED or ambulatory visits. Chief complaints are among one of the most widely-used syndromic data sources in many syndromic surveillance systems. Figure 3-2 shows some sample chief complaint records collected from a hospital.

Chief complaints as a syndromic data source present many advantages as well as challenges for public health monitoring. Chief complaint records are routinely generated and become available typically on the same day the patient is seen. As a comparison, diagnostic data typically take a much longer time to be coded and transmitted due to various logistical and infrastructural issues and the lack of IT personnel at smaller hospitals (Travers et al., 2006). Chief complaint records are typically accessible in an electronic format. The wide availability and timeliness make chief complaints an ideal syndromic data source. However, as each chief complaint entry is a concise statement often in short free-text phrases that often contain misspellings and abbreviations, cleaning chief complaint data and mapping them into more meaningful representations are typically necessary before the analytical processes take place. In Chapter 4 we will further elaborate this problem as to processing chief complaints for syndromic surveillance.

date	MEDREC	AGE	SEX	RACE	ETHNIC	chief_complaint
09/01/2004	MA116315	78	F	H	1	OTH PULMON EMBOLISM
09/01/2004	MA216315	2	M	B	2	WHEEZING
09/01/2004	MA316315	15	M	W	2	SOB
09/01/2004	MA416315	75	M	W	2	DYSPNEA

Figure 3-2. Sample chief complaint records sheet.

OTC medication sales and prescription data are indicative of certain illness (e.g., influenza), which could be timelier than patient visits, as people may visit a drug store before considering seeing a physician. However, getting additional information about the purchasers such as demographical information is often not possible. ESSENCE and EARS are among the systems that utilize OTC sales data for surveillance purpose. The RODS laboratory has built the National Retail Data Monitor (NRDM) to monitor the sales of OTC medications as a public health surveillance tool. Thousands of retail pharmacies, groceries, and mass merchandise operations have participated in the program, where the data and analytical results are made accessible to public health officials across the nation.

School or work absenteeism reported by schools and workplaces can also be used as an indicator of public health status. As no disease characterization available with the absenteeism report, school or work absenteeism data have relatively limited use in syndromic surveillance. Systems (such as EARS, ESSENCE) monitor the school or work absenteeism data as a rough-cut early indication to generate alarms that "something might be wrong" instead of telling "what is going wrong."

Highly reliable disease diagnostic data are available as part of hospital admission record when hospitalization takes place. However, there could be 1–3 days between a patient's first healthcare visit and his or her possible hospitalization, making such data less timely than many other data types. The Hospital Admission Syndromic Surveillance (HASS) system implemented at Connecticut Department of Public Health utilizes hospital admission data for syndromic surveillance.

Triage nurse calls, 911 calls, and ambulance dispatch calls also have the potential of signaling possible events and changes in the public health status. Although the phone call data are relatively timely, information concerning symptoms or signs recorded during patient calls when the patient consults healthcare providers needs to be cleaned and extracted for the use of disease characterization. NHS Direct in the UK has been used for spatiotemporal analyses to initiate prospective geographical surveillance of influenza in England (Meynard et al., 2008), based on calls about fever and vomiting placed to a national telehealth system.

International Classification of Diseases 9th edition (ICD-9) codes and International Classification of Diseases, 9th edition, Clinical Modification (ICD-9-CM) codes assigned for diagnoses and procedures are often available in today's healthcare information systems used for billing or third-party insurance reimbursement purposes. ICD-9/ICD-9-CM codes are used as a syndromic data source in many systems because of their wide availability in an electronic format. Other data sources such as laboratory test orders and results, or even news reports, are also studied by researchers as feasible early public health indicators. For instance, researchers have studied how the mass media covered disease outbreaks and the media activity affected antiviral sales as monitored by syndromic surveillance techniques (Racer, 2007). Web-accessible information sources regarding infectious diseases such as discussion forums, mailing lists, and government Web sites, and news outlets have been found valuable in early public health event detection. As the rapid growth of Internet use and wide adoption of real-time online communication continues, more and more current, highly local information about outbreaks is available and accessible by Web crawling to support situational awareness (Brownstein et al., 2008a). Researchers also propose to monitor blogs, discussion sites, and listservs to complement news coverage and the use of click-stream data and individual search queries is also a promising new surveillance source (Eysenbach, 2006). However, because of the distributed and unstructured nature of these sources of information, monitoring public health related events through them becomes a challenge. Recently two global systems, HealthMap and Argus, were developed to provide real-time global information integration and public health status monitoring (Brownstein et al., 2008a). The systems have been discussed in previous sections, and dedicated chapters describing them can be found in Part II.

There are very few studies connecting environmental factors with public health status. Serious investigation is called for to determine whether monitoring environmental indicators can assist public health surveillance. In one such study (Babin et al., 2007), air quality measurements from the Environmental Public Health Tracking Program (EPHTP) are passed to the CDC, and the relationship between air quality and pediatric emergency department (ED) visits for asthma among DC residents are quantified over a 3-year period. Studying environmental factors could help understand background disease patterns so that unexpected fluctuations could be better detected (Zeng et al., 2008).

1.1 Comparison of Data Sources

A quantitative compilation of our research results shows that most of the syndromic surveillance systems monitor a combination of data sources from multiple sites instead of relying on a single data indicator. Out of the 56 systems numerated in Tables 2-1 through 2-5, wherein the details are known, 80% use ED chief complaints (both free text and ICD-9 coded chief complaints) as a timely public health indicator. Fifty percent of the systems monitor OTC drug sales. Thirty percent of the systems use hospital admission data as one of the inputs. Thirty of the systems also collect school/work absenteeism data. However, absenteeism or drugs sales are never used alone. Fourteen systems also connect to poison centers or laboratories for test orders, or monitor 911 calls. Additionally, most ED visits chief complaints are in free text (90%), which suggests the importance of free text processing or natural language processing techniques for medical information processing in this area.

ISDS (International Society for Disease Surveillance) also conducted a survey of state syndromic surveillance use including 46 respondents in 2008. The following figure (Figure 3-3) shows the distribution of use of data sources by the surveyed syndromic surveillance system (Mostashari et al., 2008). The numbers largely align with our quantitative findings above.

A major concern regarding the data used in the surveillance activities is about the effectiveness and validity of their usage for illness pattern detection. To be valid in the context of syndromic surveillance, evidence is needed that a data source may have value in identifying an outbreak or biological attack. A number of studies have examined to some degree whether and how effective the data sources are, as well as a possible time lead compared with diagnosis. Magruder's study (Magruder, 2003) about using OTC data/sales as a possible early warning indicator of human diseases revealed about a 90% correlation between flu-remedy sales and physician diagnoses of acute respiratory conditions together with a 3-day lead time reported. Another study (Doroshenko et al., 2005) shows that nurse-led helpline calls can also be used for early event detection. SSIC (Syndromic Surveillance Information

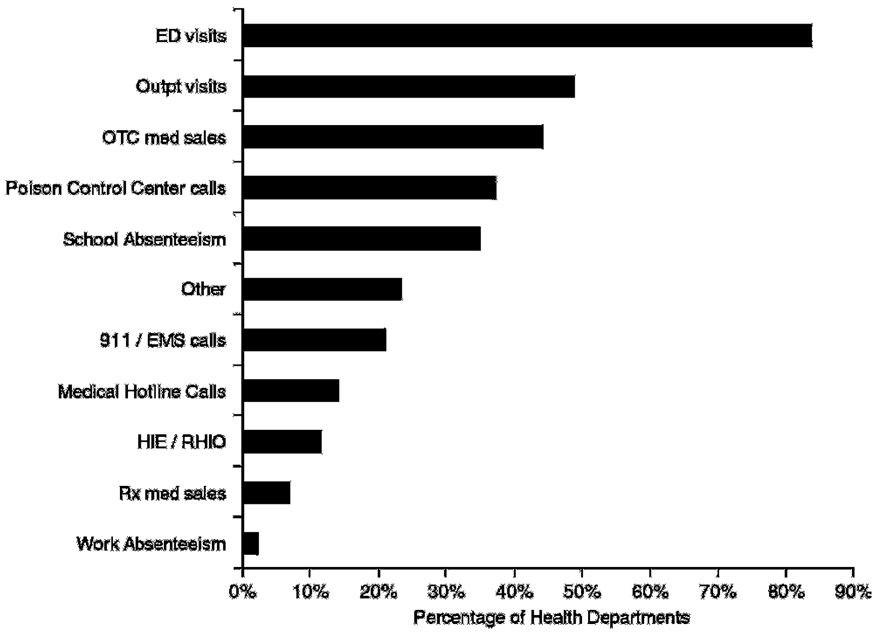

Figure 3-3. Syndromic surveillance data sources use survey by ISDS.

Collection) program tested the use of visit-level discharge diagnoses from several clinical information systems as a syndromic data source (Duchin et al., 2001; Lober et al., 2003). One limitation of using chief complaints as syndromic data is that they provide different predictive values from discharge diagnoses, as reported by (Begier et al., 2003). Generally, chief complaints best capture illnesses mainly characterized by nonspecific symptoms like fever, while discharge diagnoses appear better at tracking illnesses requiring brief ED clinical evaluation and testing, such as sepsis and possibly meningitis (Begier et al., 2003).

Although most of the syndromic surveillance systems use multiple data sources, further examination about whether the different data are telling the same story, i.e., flagging the possible outbreaks for certain illness with consistency, is necessary. Edge et al. (2004) reported correlations between OTC antinausea and antidiarrhea medication sales and ED admissions. However, in a study conducted by the Infectious Disease Surveillance Center, Japan (Ohkusa et al., 2005a), they found no evidence that sales of OTC medications used to treat the common cold correlated with influenza activities. It has been observed that as individuals may seek care in a variety of settings resulting in multiple reports for the same individual case in different data sources, combining these data sources properly presents major technical challenges due to dependencies existing among these data sources (Costa et al., 2007).

Preliminary investigations have evaluated the effectiveness of different data sources in syndromic surveillance and studied the differences among them in terms of information timeliness and characterization ability for outbreak detection, as they represent various aspects of patient healthcare-seeking behavior (Ma et al., 2005). For example, school/work absenteeism comes to notice relatively earlier as individuals take leave before seeking healthcare in hospitals or clinics, but specific disease evidence provided by the absenteeism type of data is limited. Table 3-1 provides a classification of different data sources used for syndromic surveillance organized by their timeliness and capability to characterize epidemic events.

Table 3-1. Data sources and their timeliness and disease characterization capability.

Data source	Description	Specif-icity *	Timel-iness **	Advantages	Weaknesses
Chief complaints from ED visits or ambulatory visits	Patient-reported signs and symptoms of their illness (e.g., coughing, headache, etc.) (Bradley et al., 2005; Espino and Wagner, 2001; Lombardo et al., 2004)	H	M-H	Routinely generated; available typically on the same day the patient is seen; and often available in electronic format	Available in short free-text phrases that contain mispellings and abbreviations; need to be cleaned; vocabulary differences across hospitals
ED diagnosis data	Diagnosis data available in electronic form from EDs (Travers et al., 2006)	H	L	Widely available in electronic format	Typically available a week after an encounter
OTC medication sales, prescription medication data	Medication sales data indicative of certain illness (e.g., influenza) as patients seek remedies (Besculides et al., 2004; Thomas et al., 2005)	M-H	H	Providing early signs and indications more timely than patient visits; data routinely generated and available in electronic format	Additional information about medication purchasers unknown
School or work absenteeism	Collected from school or workplace (Besculides et al., 2004; Thomas et al., 2005)	L-M	H	Timely	Lack of disease characterization (Quenel et al., 1994)
Hospital admission	Data are recorded when hospitalization takes place (Dembek et al., 2005; Dembek et al., 2004)	H	M	Highly reliable disease diagnosis	Generally an interval (1-3 days) exists between the first healthcare visit and admission (Buehler et al., 2003)

Data source	Description	Specif-icity *	Timel-iness **	Advantages	Weaknesses
Triage nurse calls, 911 calls	Symptoms of signs recorded during patient calls consulting healthcare nurses (Crubézy et al., 2005).	H	H	Relatively timely, as patients usually make phone calls before office visit	Need to be cleaned
ICD-9 (International Classification of Diseases, 9th edition) coded billing info	Preliminary diagnosis for billing (Begier et al., 2003; Espino and Wagner, 2001; Tsui et al., 2001)	H	M	May provide a better positive predictive value than chief complaints. available in most electronic medical systems	Often available after a relatively brief ED evaluation (days or weeks after an encounter)
ICD-9-CM (International Classification of Diseases, 9th edition, Clinical Modification)	Allow assignment of codes to diagnoses and procedures; often used for third-party insurance reimbursement purposes	H	M	Relatively timely and specific regarding illness characterization	Often assigned to patient visit days or even weeks after patient encounter
Laboratory test orders	Orders for laboratory tests (Wagner et al., 2001)	M	M-H	Relatively timely and specific regarding illness characterization	
Laboratory test results	Results of laboratory tests	H	L	Disease cases can be reported with high reliability	Lack of timeliness (test results may take more than a week)
Open source information (local or regional events)	Official or unofficial news reports, bulletin notification, Web forums and other online media	L	L	Tremendous amount of information that is freely available	Distributed and unstructured nature

*Disease characterization capability (Low-L, Medium-M, High-H)

**Timeliness of data to enable detection of outbreaks before confirmed diagnosis (Low-L, Medium-M, High-H)

2. STANDARDIZED VOCABULARIES

Data standard development, or more generally interoperability, is a key to successful, cross-jurisdictional syndromic surveillance. A standardized syndromic data representation would have a number of implications. First, a specialized vocabulary enables accurate representation for communicating information and events. Data formats and coding conventions that are inconsistent among different sites (e.g., laboratory tests and results can be reported in multiple ways) could be an obstacle in capturing illness cases.

More importantly, streamlining the delivery of electronic data across multiple sites saves time and eventually enables real-time reporting and alerting. Real-time data transmission and event reporting with a universal data format standard and messaging protocol is a primary motivator in the development of syndromic surveillance systems. Because of the varying internal data structures and database schema among various healthcare information systems, it takes a significant amount of time and processing resources for data conversion and normalization. According to an estimation in 2004, the use of data exchange standards in healthcare could save up to $78 billion annually (Pan, 2004).

In addition, syndromic surveillance systems that are more complex and geographically distributed need to be interoperable to enhance jurisdictional collaboration for timely event detection and response. Therefore, developing and imposing standards from programmatic, constructive, architectural, and managerial perspectives is especially addressed by the CDC-led syndromic surveillance initiatives. These initiatives are a collaborative effort involving the Public Health Information Network (PHIN) framework (CDC, 2006c), the National Electronic Disease Surveillance System (NEDSS) (CDC, 2004), the National Center for Vital Health Statistics, Department of Defense, Department of Veteran Affairs, and all National Institutes of Health.

This section discusses the development, adoption, and implementation of standard vocabularies for electronic emergency room records, laboratory testing, clinical observations, and prescriptions, along with the messaging standard to transport these records. Many available code standards currently used in syndromic surveillance have been borrowed from public health systems (Wurtz, 2004). Current efforts to standardize vocabulary are based on Logical Observation Identifiers Names and Codes (LOINC®), Systematized Nomenclature of Medicine (SNOMED®), International Classification of Diseases, Ninth Revision (ICD-9), and Current Procedural Terminology (CPT®) as core vocabularies. In addition, Unified Medical Language System (UMLS) has been used as cross reference ontology among the above coding systems. Health Level Seven (HL7) is used as a messaging standard in public health.

2.1 Existing Data Standards Used in Syndromic Surveillance

Here we provide a brief summary of each coding system to illustrate their scope and target medical domain.

UMLS: The Unified Medical Language System (UMLS) (Fung et al., 2006) provides a cross reference ontology among a number of different biomedical coding systems and standards, and a semantic structure defining relationships among different clinical entities. Its Semantic Network and Metathesaurus help facilitate system developers in building or enhancing electronic information systems that integrate and/or aggregate biomedical and health data and knowledge.

LOINC: LOINC codes are universal identifiers for laboratory and other clinical observations. Distinct LOINC codes are assigned based on specimen types (e.g., “ser” = serum) and methods of the test (e.g., immune fluorescence), with specific description for different conditions. As LOINC codes were originally developed for billing purposes, they do not convey information about the purpose or results of the test (Wurtz, 2004). The CDC has developed “Nationally Notifiable Conditions Mapping Tables” (http://www.cdc.gov/PHIN/data_models), which provide mappings from LOINC codes to nationally-notifiable (and some state notifiable) diseases or conditions.

SNOMED: SNOMED is a nomenclature classification scheme for indexing medical vocabulary, including signs, symptoms, diagnoses, and procedures. It defines code standards in a variety of clinical areas called coding axes. It can identify procedures and possible answers to clinical questions that are coded through LOINC.

ICD-9-CM: ICD-9-CM was developed to allow assignment of codes to diagnoses and procedures associated with hospital utilization in the United States and are often used for third-party insurance reimbursement purposes. Table 3-2 shows a partial code set used by ESSENCE for fever.

Table 3-2. ICD-9-CM coding examples.

ICD9CM	ICD9DESCR
020.2	PLAGUE, SEPTICEMIC
020.8	OTHER TYPES OF PLAGUE
020.9	PLAGUE NOS
021.8	TULAREMIA NEC
021.9	TULAREMIA NOS

An updated release of ICD-10-CM was made available in 2007 for public viewing. The codes of ICD-10-CM are now under testing and not currently valid for any purpose or use. A research study has been conducted to examine the usefulness of the ICD-10-CM system in capturing public health diseases, when compared with ICD-9-CM. The study also examined agreement levels of coders when coding public health diseases in both ICD-10-CM and ICD-9-CM. Overall results demonstrate that ICD-10-CM is more specific and captures more of the public health diseases examined than ICD-9-CM (Watzlaf et al., 2007).

HL7: HL7 (HL7, 2006; Hooda et al., 2004; Thomas and Mead, 2005) is the ANSI-accredited healthcare standard messaging format, used for transmitting information across a variety of clinical and administrative healthcare information systems. It specifies the syntax that describes where a computer algorithm can find various data elements in a transmitted message, enabling it to parse the message and reliably extract the data elements contained therein. HL7 Version 2.3 provides a protocol that enables the flow of data between systems. HL7 Version 3.0 (Beeler, 1998) is being developed through the use of a formalized methodology involving the creation of a Reference Information Model to encompass the ability not only to move data, but also to use it once it has been moved.

Development and adaptation of coding standards and standardized messaging formats are essential for information exchange and sharing, a prerequisite for public health surveillance. However, different standards and implementations exist for operational clinical, laboratory, and hospital information systems, which causes significant obstacles for information sharing. Nonetheless, standards are being developed, improved, and adopted increasingly widely.

Table 3-3. Adopted healthcare information standards in syndromic surveillance.

Clinical vocabulary	Main contents	Advantages	Limitations
UMLS	The UMLS Metathesaurus is a collection of different source vocabularies, organized according to meaning and lexical characteristics of terms. The Semantic Network contains explicit biomedical concepts and relationships.	Provides cross referencing between multiple vocabularies	Lacking granularity for medical diagnosis and syndromic surveillance (Lu et al., 2006)
LOINC	Laboratory results and observations. Could refer to a laboratory value (e.g., potassium, white blood cell count) or a clinical finding (e.g., blood pressure, EKG pattern)	Contains many genetic tests. It is mapped to UMLS and SNOMED RT and CT	Not suitable to capture the purpose or results of the test
SNOMED-CT (SNOMED-Clinical Terminology)	Used to distinguish concepts for the condition (e.g., pertussis) and the causative organism (e.g., *Bordetella pertussis*), suitable to code laboratory results, nonlaboratory interventions and procedures, and anatomy and diagnosis	Combines SNOMED RT and Clinical Terms Version 3	Proprietary
SNOMED-RT (SNOMED-reference terminology)	Includes concepts and terms for findings (disorders and clinical findings by site, method, and function), normal structures (anatomy/topography) and abnormal structures (pathology/morphology)	Well-tested and used in the field for decades	Proprietary
ICD-9-CM	Used to code morbidity data, final diagnosis, procedures, and reimbursement	Widely used (state-mandated)	Not suitable for clinical documentation of diagnoses, symptoms, signs and problem lists. (Hogan et al., 2002)
ICD-10-CM	ICD-10-CM represents a significant improvement over ICD-9-CM. Some specific improvements include: the addition of information relevant to ambulatory and managed care encounters; expanded injury codes; the creation of combination diagnosis/symptom codes to reduce the number of codes needed to fully describe a condition. It allows greater specificity in code assignment(2008)	Not deployed yet	Testing studies have demonstrated that ICD-10-CM is more specific and fully captures more of the public health diseases examined than ICD-9-CM. (Watzlaf et al., 2007)

In addition to leveraging existing healthcare standards, some groups have proposed additional coding and messaging standards tailored specifically for syndromic surveillance. For example, the Frontlines group (Barthell et al., 2002, 2004) is focusing on the development of standard reporting and coding structures specific to syndromic data. They defined the data elements in triage surveillance reports and a set of codified values for chief complaints. They also proposed a system to facilitate continuous flow of XML-based triage report data among hospital EDs, and state and local health agencies. The ongoing effort motivated to develop an electronic health record is largely relevant as well to public health surveillance from the point of view of coding and messaging standards. For instance, the Veterans Administration (VA) has been standardizing its clinical terminology to comply with industry-wide standards. In the National VA Health Data Repository (HDR), "Unique enterprise identifiers are assigned to each standard term, and a rich network of semantic relationships makes the resulting data not only recognizable, but also highly computable and reusable in a variety of applications, including decision support and data sharing with partners such as the Department of Defense (DoD)" (Bouhaddou, Lincoln et al., 2006).

In addition to technical considerations, regulatory and compliance issues also need to be examined carefully to address data standardization challenges. For instance, the US has implemented laws, such as HIPAA's Administrative Simplification, to enforce standardization in healthcare information by mandating, for example, health plans, healthcare clearinghouses, and providers that conduct certain transactions electronically comply with the HIPAA transaction standards.

Despite the availability of standard vocabularies discussed above, healthcare providers and public health researchers and practitioners often use natural language when describing biomedical concepts and constructs, even in the context of highly structured case report forms. Hunscher et al. (2006) described work in progress and lessons learned in translating complex natural-language concepts on case report forms into machine-readable format using the HL7 CDA, LOINC, and SNOMED-CT standards.

3. DATA ENTRY AND DATA TRANSMISSION

Syndromic data are being collected through various kinds of healthcare and public health information systems. Such data collection efforts often have to cross organizational boundaries and jurisdictions. This section discusses related data entry and transmission techniques.

3.1 Data Entry Approaches

Data entry approaches for syndromic surveillance fall into four categories: paper-based forms, Web-based interface, local data input software application, and hand-held devices (Zelicoff et al., 2001). Many systems support multiple data entry approaches as they involve multiple sites with possibly different IT infrastructure support (Espino et al., 2004; Lombardo et al., 2003). In general, the manual approach using paper-based forms can lead to unwanted delays as the records have to be converted later to an electronic format.

3.2 Secure Data Transmission

Secure data transmission is critical to data integrity and confidentiality. The specific challenges are as follows. How can a syndromic surveillance system retrieve syndromic data from data providers (e.g., hospitals and pharmacies)? How can data transfers be done securely over the communication channels such as the Internet?

The existing transmission approaches are either automated or manual. Automated transmission refers to transferring of data over a communication media where human intervention (e.g., to initiate each transmission transaction) is not required. Manual transmission entails significant human intervention. About 33% of the 50 systems surveyed rely primarily on automated data transmission, whereas the remaining 67% rely on human intervention in both data requesting and receiving. Email messages with text reports or data files as attachments, despite the security and data exposure risks, are still widely used to transfer syndromic data from clinical systems to syndromic surveillance systems.

The XML-based HL7 messaging standards play an important role in automated data transmission, since a significant portion of health systems support HL7. Among the systems surveyed, those capable of automated data transmission all use HL7 one way or another. For example, the RODS system and the BioPortal system use HL7 messaging protocols for automatic syndromic data transmission. In RODS, an HL7 listener implemented as Enterprise JavaBean (EJB) beans is used to receive HL7 messages from each underlying health system. The messages transmitted are first parsed by an HL7 parser bean before being loaded into the database. A configuration file written in XML is used to specify the hierarchical structure of the data elements in each HL7 message (Tsui et al., 2003). BioPortal also relies on an HL7-based approach to transmit data as HL7-compliant XML messages. This approach allows for dynamic changes in the message structure (Hu et al., 2005; Zeng et al., 2004b).

Compared with other approaches that mainly support file-based transmissions in a batch mode, HL7-based approaches are more efficient and effective. According to a RODS study (Tsui et al., 2005), they could reduce reporting latency by 20 hours. Secure networking techniques such as VPNs (Virtual Private Networks), SSL (secure socket layer), HTTPS, and SFTP (secure file transfer protocol) are now being increasingly utilized (Rhodes and Kailar, 2005).

Is there a best approach to transmit data from data providers to syndromic surveillance systems and the involved public health agencies? There is no simple answer to this question. Typically the IT infrastructure of the data providers (e.g., hospitals) needs to be upgraded to enable timely, reliable, and secure data collection.

Many practical challenges hindering the data collection effort also need to be addressed, including: (1) providing and transmitting data either requires staff intervention or dedicated network infrastructure, which often require extra costs; (2) data sharing and transmission must comply with HIPAA and other privacy regulations; (3) reducing data acquisition latency has important implications to syndromic surveillance yet is difficult and can be costly; (4) data quality concerns (e.g., incompleteness and duplications) often pose additional challenges. In particular, data ownership, confidentiality, security, and other legal and policy-related issues need to be closely examined. When infectious disease datasets are shared across jurisdictions, important access control and security issues should be resolved in advance between the involved data providers and users (Hu et al., 2005).