Technical Description of the Distribute Project: A Community-based Syndromic Surveillance System Implementation

William B. Lober[1], Blaine Reeder[2], Ian Painter[1], Debra Revere[1], Kim Goldov[1], Paul F. Bugni[1] Justin McReynolds[1], Donald R. Olson[3]

1. University of Washington, Seattle, WA
2. University of Colorado | Anschutz Medical Campus, Aurora, CO
3. International Society for Disease Surveillance, New York City, NY

Abstract

This paper describes the design of a syndromic surveillance system implemented for community-based monitoring of influenza-like illness. The system began as collaboration between colleagues from state and large metropolitan area health jurisdictions, academic institutions, and the non-profit, International Society for Disease Surveillance. Over the six influenza seasons from 2006 to 2012, the system was automated and enhanced, with new features and infrastructure, and the resulting, reliable, enterprise grade system supported peer comparisons between 44 state and local public health jurisdictions who voluntarily contributed summarized data on influenza-like illness and gastrointestinal syndromes. The system was unusual in that it addressed the needs of a widely distributed, voluntary, community engaged in real-time data integration to support operational public health.

Keywords: syndromic surveillance, secondary use of health data, Internet, public health standards, surveillance practice

Introduction and Overview

In this paper we describe the technical components, architecture and processes of the Distribute system implementation as developed for the Distributed Surveillance Taskforce for Real-time Influenza Burden Tracking & Evaluation (DiSTRIBuTE) Project. Initiated in 2006 by the International Society for Disease Surveillance (ISDS), and operated by ISDS with technical support from the University of Washington (UW), the project enabled community-based public health syndromic surveillance.

Syndromic surveillance is the practice of monitoring existing health data sources for early detection and ongoing monitoring of adverse changes in population health [1]. The objective of public health syndromic surveillance is to collect, analyze, and interpret data about health events to achieve early detection of an event of public health interest such as an influenza outbreak and

to provide timely dissemination of collected information to decision makers. Monitoring for influenza outbreaks is of particular value for syndromic surveillance [2]. This monitoring, typically based on influenza-like illness (ILI) chief complaint data extracted from emergency department (ED) or urgent care (UC) facility electronic medical records (EMRs), can provide advance warning of an influenza season before laboratories can confirm results of viral isolates [3]. Chief complaints from patients that indicate ILI include symptoms such as "fever", "cough" and "sore throat".

Greater detail regarding the history, epidemiologic basis, and community participatory nature of the Distribute project have been published previously [4-6]. Briefly, the Distribute project began as a set of methods and procedures developed by the New York City Department of Mental Health and Hygiene to monitor age-stratified ED visits for fever and respiratory illness, based on data from the 2001-2001 and 2005-2006 influenza seasons [7]. The Distribute project was initiated in 2006, and during the 2007-08 influenza seasons, the system was operated using both manual and automated processes to integrate and visualize data from eight participating state and local health jurisdictions [5,6]. Findings from this surveillance activity were disseminated via a public web site, www.syndromic.org.

By 2011, 44 state and local public health departments, from 29 states, transmitted summarized daily ED and ambulatory care visit data using locally defined ILI syndromes to Distribute. Six sites also transmit data on visits for illnesses classified as gastrointestinal (GI) syndrome. Data sent to the Distribute system implementation were validated and normalized before being aggregated into a common data repository. Metadata were maintained within the Distribute system implementation to provide context for the surveillance data. These metadata gave information about the contributing health jurisdictions, about the methods they use to collect their surveillance data locally, including definitions for different syndrome categories, and performance characteristics of the data transmission from the jurisdiction. Both data and metadata were maintained through automated and manual curation processes. Aggregated syndromic surveillance data were made available to member jurisdictions and other stakeholders through visualizations, downloads on a private web site and APIs. One set of APIs was used by ISDS to maintain a public Web site showing a limited view of weekly data in the Distribute system implementation, at www.isdsdistribute.org.

Three perspectives demonstrate signature characteristics of the Distribute system implementation in 2012, and these characteristics were reflected in the system's architecture. From an organizational perspective, the Distribute project was a voluntary system. State and local health jurisdictions elect to participate as a way of comparing their data with that of neighboring jurisdictions. From an epidemiologic perspective, the Distribute system implementation relied on summarized ED/ambulatory care visit counts and totals, not visit level data, or so-called "line listings". And, from a technical perspective—the focus of this paper—the system underwent a transition from manual processes used to monitor the 2007-08 influenza season [4-6] to the fully automated processes and infrastructure described herein, which have evolved through the subsequent years of seasonal and pandemic (H1N1) influenza.

Technical Overview

Figure 1 displays an overview of the process flow of services of the Distribute system implementation architecture, grouped as Authentication, Upload, Notification, Application and Data Quality Services. All services rely on a central database. When ILI data files are uploaded to the system, Notification Services send a confirmation e-mail to both the uploader and UW CIRG technical staff. Application Services route the data to a central database repository where the data are run through a series of automated data quality (DQ) processes. A DQ team manually reviews data daily and weekly to detect anomalies. The development (operations) team is on call to handle any aberrations in system function. Once data are processed, they are made available in 2 ways: the Distribute project public Web site (http://www.isdsdistribute.org/) which is open to anyone via the Internet and the Distribute system implementation restricted Web site which requires a member ID. Components of the Distribute system implementation are described in detail in the sections that follow.
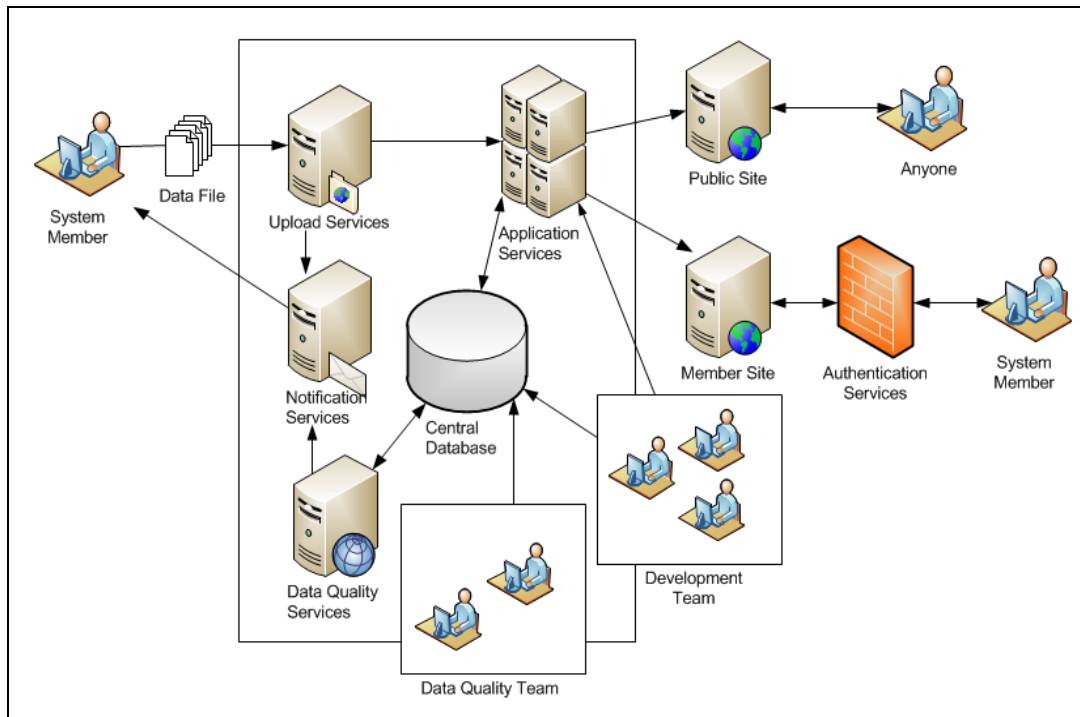

Figure 1. Overview of the process flow in the Distribute system implementation.

System Architecture

Hardware.

The Distribute system implementation runs in a three-tiered development environment: a development server, production and staging server, and an upload and metadata-editing server. Three virtual machines host Apache servers and MySQL databases running on Linux. The central database repository consists of a master database and two slave databases where data are replicated for fault tolerance and to ensure data access for uploads, development and Application

Services. The central database embodies a star schema--well known for data warehouse design [8] --that relies on several dimension tables to describe stratified or distinct observations and facilitates flexibility in managing multiple definitions for the same data set. For example, age groups can be defined on-the-fly versus being hard coded and user-defined syndromes can be accommodated with a minimum of effort. A scheduled periodical scheduled job improves query performance for common elements by generating aggregated tables that contain a smaller number of stratifying elements (jurisdiction ID, ED visit date) and synchronizes the master database with the two slave databases to improve performance and support scaling. The system is implemented as a series of integrated virtual machines (VM) on a KVM framework [http://www.linux-kvm.org]. The complete national implementation is distributed among 3 VMs hosted on a single commercial-grade server (16 cores, 32 GB, 400 GB RAID 10). Additional server hardware in alternate locations is available to support high reliability. The VM strategy also supports deployments on commercial "cloud" infrastructure, such as Amazon's Elastic Computing Cloud (EC2).

Authentication.

To accommodate convenient login to the restricted Web site (see Fig. 2), UW CIRG developed and implemented a novel Authentication Services package, developed as a module for the open source Apache2 web server, and called "Apache2::AuthAny" (AuthAny) which is part of a codebase that will be open-sourced. AuthAny supports authentication through Google, basic authentication, Shibboleth (UW and Protect Network), and LDAP (Lightweight Directory Access Protocol) by linking different accounts into a single identity
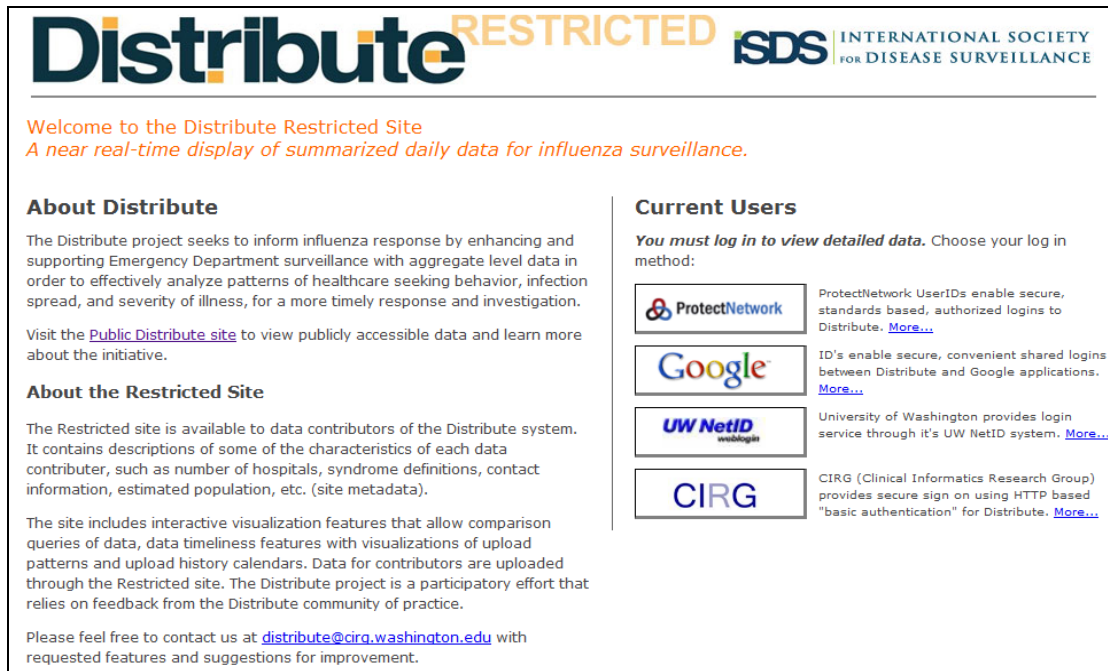

Figure 2. Login page for the Distribute system implementation restricted Web site

Uploads.

New data files can be uploaded manually or automatically via HTTP or SFTP protocols. Typically, jurisdictions providing data use automated extraction processes and automated uploading locally. These files are added to a queue for processing where they are picked up by data transformation and import processes from Application Services. Data formats currently supported include: 1) a data aggregation file in comma separated value (CSV) format; 2) a CSV file with GIPSE-formatted [9] data provided by CDC for BioSense [10] data uploads; and 3) a file with GIPSE-formatted XML data provided by a regional Health Information Exchange. These file formats were developed and adopted as necessary to aggregate data from data-providing jurisdictions and to reduce burden for meeting rigid data formatting requirements. While most participating jurisdictions upload files with data for a single jurisdiction, CDC uploads files with data for multiple jurisdictions. Data aggregation files contain separate broad and narrow ILI syndrome definitions.

Application *Services*.

Application Services for the Distribute system implementation are protected by a restricted Web site that requires AuthAny authentication. Here members can view metadata about each data-providing jurisdiction, see data timeliness trends and run interactive visualizations that show comparison of ILI trends for different regions. Application Services are made possible through import and data transformation processes, two application programming interfaces (APIs) and visualization, monitoring and editing tools. When the periodic job for Upload Services detects a new file it: 1) identifies the file format by checking the metadata for the uploading jurisdiction; 2) calls the transformer process appropriate to the data file type for import and storage to the central database; 3) validates the date format and syndrome definitions of the transformed data; and 4) imports data that are syntactically valid to the central database. The data transformation process generates a human readable report that is logged and displayed in an Events log (see Fig. 3) along with messages about successful uploads.

| Home | Site Profile | Events |

Events

Show/Hide by Type (click to change) (only showing most recent 1000)
parse: off   login: off   import: on   **upload: on**   delete: on   **account: on**   **project: on**

Time	Duration	Type	Site	User	Description
2011-03-17 16:43:01	-	upload	somesite	someuser	someuser uploads somesite_03172011.csv via ftp for site
2011-03-17 16:43:01	-	upload	somesite	someuser	someuser uploads somesite_03172011.csv via ftp for site
2011-03-17 16:00:23	39 sec	import	somesite		updating compound data
2011-03-17 16:00:22	1 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:20	2 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:20	1 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:17	3 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:16	1 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:16	1 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:16	1 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:07	9 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:05	2 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:03	2 sec	import	somesite		importing somesite_03172011.csv
2011-03-17 16:00:01	-	upload	somesite	someuser	someuser uploads somesite_03172011.csv via ftp for site
2011-03-17 16:00:01	-	upload	somesite	someuser	someuser uploads somesite_03172011.csv via ftp for site

Figure 3. Events log in the Distribute implementation system restricted Web site.

APIs.

The Distribute system implementation features two APIs: a Data API and a Metadata API. The Data API is used internally to allow authenticated applications to access specific types of data in the system to monitor DQ, create visualizations and make comparisons of regional data sets and by the public Distribute site to visualize data. The Metadata API is designed to work in conjunction with the Data API. It is used internally to access the metadata for participating state and local health jurisdictions and by the public Distribute site to access background and event information

Interacting with Syndromic Surveillance Data

The Distribute system implementation generates three types of visualizations on the Web site: interactive "annotated time line" visualizations, time-series graphs, and heat maps. The "annotated time line" graphs use a Google JavaScript library to generate an Adobe Flash [11] widget based on Distribute Data API. "Time Series Graphs" are PNG images created by the Google Chart API in response to an HTTP request (see Fig 4 example comparing ILI broad and ILI narrow syndromic definitions for a de-identified jurisdiction). These images are generated using by scripts that call our Data API, encode the output into query parameters, issue GET requests to Google, and save the resultant PNG files on our web server. Heat map visualizations are also periodically generated and cached on the server, using code that integrates calls to the open source "R" statistical software.
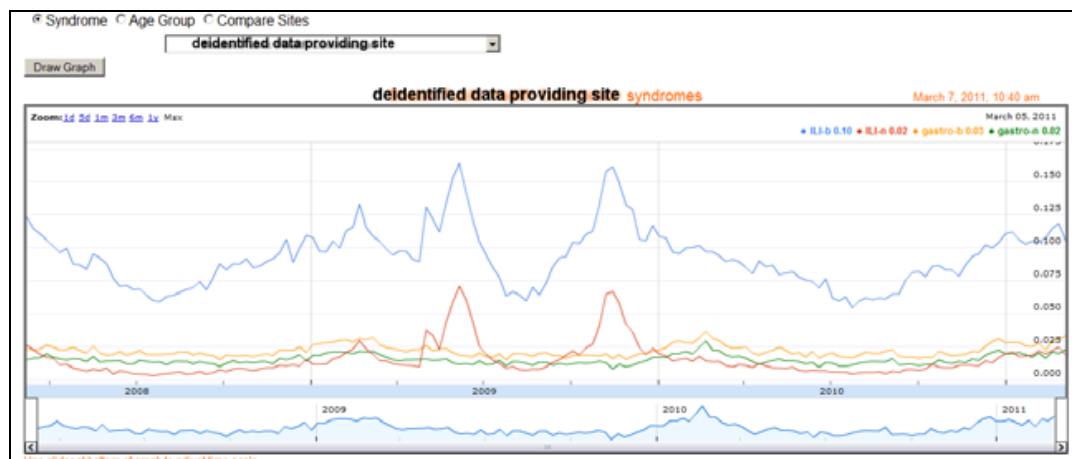

Figure 4. View of an ILI broad and ILI narrow comparison for a de-identified jurisdiction in Distribute system implementation Web site.

Jurisdiction Metadata.

Metadata provide the context and background of the data provided by each contributing jurisdiction. A metadata editing tool (see Fig. 5) is allows administration and modification of metadata content (for example, number of health care facilities from which data are collected; free-text descriptions of ILI syndrome definitions; contact information for each jurisdiction, etc).

6

All fields below are optional - fill in only those which are pertinent to this site, and be sure to hit save.

[Save Changes]

Syndrome Description(s):	ILI-b is looking for febrile illness (includes a search for fever, high temp, elevated temp, shivers, feeling hot, etc. as well as a search for the ICD-coded diagnosis of fever). ILI-n searches through chief complaints and diagnoses for mention of either fever+cough, fever+ST, or influenza.
	i.e. how do you define ILI-broad and ILI-narrow?
How Are The Data Stratified:	age groups, zip3 (patient)
	i.e. by age groups, temperature ranges, disposition, patient Zip3, institution Zip3, etc.
Types of visits included:	ED only
	e.g. EDs only, includes primary care and inpatient
Number of Clinical Sites Reporting:	18 EDs

Figure 5. Metadata editor in the Distribute system implementation restricted Web site.

Another type of metadata is concerned with the behavior of the data, such as upload patterns. Fig. 6 is a de-identified visualization of uploads for a number of sites over the last 100 days . A long blue bar indicates a site has uploaded every day. White gaps indicate interruptions to the uploads, and make evident the pattern of activity. In addition to daily and weekly manual data inspection, these views can help determine if there are upload anomalies based on usual upload behavior.

Figure 6. Data timeliness view for uploads in the Distribute system implementation restricted Web site.

Public site.

Data from the system is accessed through an API to produce a site and set of visualizations for public access based on weekly trends for a trailing one-year period, [www.isdsdistribute.org]. That site was developed and is maintained for ISDS by programmers working at the Boston Childrens' Hospital Informatics Program. National trends can be seen through a map display of aggregated data from the ten Federal health regions, and site visitors may "drill down" to see the recent trends in the localities of data-providing jurisdictions that allow this broad access to their data.

Gossamer Health Platform

The Distribute system implementation runs on the Gossamer Health (Good Open Standards System for Aggregating, Monitoring and Reporting Health) platform, an open source framework developed by the UW Clinical Informatics Research Group (CIRG). Work on Gossamer Health began in 2001 with the Syndromic Surveillance Information Collection (SSIC) project [12,13] and has continued in the development of novel informatics applications and approaches to data collection, integration, storage, aggregation, display, and dissemination, in several domains, including surveillance and clinical research support [12-18]. The Distribute system implementation has contributed important advances in notification, authentication, and visualization to the Gossamer Health platform.

Innovation and Customization

The Distribute system implementation has been tailored for specific uses in two different ways. Some applications of the Distribute system implementation have used new syndromes while other implementations of the system have created purpose- specific systems with distinct sets of data. Six sites are presently providing gastro-intestinal data as a way of exploring the utility of the system for the comparison of public health indicators beyond influenza, using the same community-based model. Similarly, a research collaborative within the community of practice uses the Distribute system implementation to investigate the utility of a common syndrome across jurisdictions. While at first it seems obvious that a single syndrome definition would make data from multiple jurisdictions more comparable, in fact local variation in health care systems, coding practices, and clinical practice creates variation in the underlying clinical data which may be compensated for by local variation in syndrome definition and "tuning" by public health agencies. It remains an open question whether the best national or regional picture of surveillance comes from combining syndrome data based on a uniform definition, or syndrome definition data based on a "best local" definition. While this question is well outside the scope of this paper, it has been useful for the Distribute system implementation to develop the technical capabilities to support this type of inquiry.

Purpose-specific systems have been created both for demonstration and operational uses. Since 2009, the Distribute system implementation has been used as part of the Integrating the Healthcare Enterprise (IHE) Showcase demonstration of a model health information exchange (HIE) at the annual Health Information Management Systems Society (HIMSS) meeting, and that the 2009 Public Health Information Network meeting. In those settings, the system has been used to demonstrate the surveillance role of state and federal public health agencies in demonstrations of standards-based interoperability. The data in those systems was typically derived from real-time monitoring of simulated patient events within the Showcase. The Distribute system implementation was also used for an operational purpose-specific system, by making aggregate, de-identified syndrome data from states and sub-state regions in the Pacific Northwest available to Canadian agencies immediately before and during the 2010 Winter Olympics in Vancouver, British Columbia.

Conclusion and Future Work

The Distribute system implementation has served its purpose in demonstrating the value of a community-based system in facilitating local and state public health agencies access to the surveillance indicators of peer jurisdictions. While parts of this architecture are somewhat routine, and other parts are somewhat novel, the Distribute system implementation has given us a sense of the complexity required for a system that serves the need for reliable, elective sharing of a relatively homogeneous set of data among a community of voluntary contributors.

Going forward, the data handling and visualization capabilities of the system are well aligned with the public health business processes identified by ISDS as part of their recommendations for the syndromic surveillance meaningful use criteria [19]. The Distribute system implementation has the ability to incorporate multiple indicators, and to serve as a comparison platform across jurisdictions, are important basic capabilities, but it is important for the public health community to develop a roadmap for their needs to drive the further development of the system

Acknowledgements

References

1. Henning K. 2004. Overview of Syndromic Surveillance What is Syndromic Surveillance? *MMWR Morb Mortal Wkly Rep*. 53(Suppl), 7-11.

2. Buehler J, Sonricker A, Paladini M, Soper P, Mostashari F. 2008. Syndromic Surveillance Practice in the United States: Findings from a Survey of State, Territorial, and Selected Local Health Departments. *Advances in Disease Surveillance*. 6(3), 20.

3. Heffernan R, Mostashari F, Das D. 2004. Syndromic Surveillance in Public Health Practice, New York City. *Emerg Infect Dis*. 10(5), 858-64. and. PubMed http://dx.doi.org/10.3201/eid1005.030646

4. Olson DR, Paladini M, Buehler J, Mostashari F. 2008. Review of the ISDS Distributed Surveillance Taskforce for Real-time Influenza Burden Tracking & Evaluation (DiSTRIBuTE) Project 2007/08 Influenza Season Proof-of-concept Phase. *Advances in Disease Surveillance*. 5(3), 185.

5. Diamond CC, Mostashari F, Shirky C. 2009. Collecting And Sharing Data For Population Health: A New Paradigm. *Health Aff (Millwood)*. 28(2), 454. and. PubMed http://dx.doi.org/10.1377/hlthaff.28.2.454

6. Olson DR, Paladini M, Lober WB, Buckeridge DL. ISDS DiSTRIBuTE Working Group. Applying a New Model for Sharing Population Health Data to National Syndromic Influenza Surveillance: DiSTRIBuTE Project Proof of Concept, 2006 to 2009. *PLoS Currents: Influenza*. 2011;3:RRN1251 PubMed http://dx.doi.org/10.1371/currents.RRN1251

7. Olson DR, Heffernan RT, Paladini M, Konty K, Weiss D, et al. 2007. Monitoring the Impact of Influenza by Age: Emergency Department Fever and Respiratory Complaint Surveillance in New York City. *PLoS Med*. 4(8), e247. and. PubMed http://dx.doi.org/10.1371/journal.pmed.0040247

8. Krippendorf M, Il-Yeol S, eds. The translation of star schema into entity-relationship diagrams. Database and Expert Systems Applications, 1997 Proceedings, Eighth International Workshop on; 1997 1-2 Sep 1997.

9. Dixon B. Geocoded Interoperable Population Summary Exchange (GIPSE) v1.0 Nationwide Health Information Network (NHIN) 2009.

10. Bradley CA, Rolka H, Walker D, Loonsk J. BioSense: Implementation of a National Early Event Detection and Situational Awareness System. IMMW MMWR Morbidity & Mortality Weekly Report2005. p. 11-9.

11. Adobe Systems Incorporated. Adobe Flash Platform. Adobe Systems Incorporated; 2011 [cited 2011 3/17/2011]; Available from: http://www.adobe.com/flashplatform/.

12. Duchin JS, Karras BT, Trigg LJ, Bliss D, Vo D, Ciliberti J, et al. Syndromic Surveillance for Bioterrorism Using Computerized Discharge Diagnosis Databases. 2001.

13. Lober WB, Trigg LJ, Karras BT, Bliss D, Ciliberti J, et al. 2003. Syndromic Surveillance Using Automated Collection of Computerized Discharge Diagnoses. Journal of Urban Health. *Bull N Y Acad Med*. 80, i97-106.

14. Travers DA, Waller A, Haas SW, Lober WB, Beard C. Emergency Department data for bioterrorism surveillance: electronic data availability, timeliness, sources and standards. AMIA Annual Symposium proceedings / AMIA Symposium 2003:664-8.

15. Lober WB, Trigg LKB. Information System Architectures for Syndromic Surveillance. IMMW MMWR Morbidity & Mortality Weekly Report 2004:203-8.

16. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, et al. 2004. Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience. *Journal of the American Medical Informatics Association: JAMIA*. 11(2), 141. and. PubMed http://dx.doi.org/10.1197/jamia.M1356

17. Hills R, Lober W, Painter I. Biosurveillance, Case Reporting, and Decision Support: Public Health Interactions with a Health Information Exchange. In: Zeng D, Chen H, Rolka H, editors. Biosurveillance and Biosecurity: Springer Berlin / Heidelberg; 2008. p. 10-21.

18. Painter IS, Hills RA, Lober WB, Randels KM, Sibley J, et al. 2008. Extending functionality of and demonstrating integrated surveillance for public health within a prototype regional health information exchange. *AMIA Annu Symp Proc*. 6(969), 969. . PubMed

19. International Society for Disease Surveillance. Final Recommendation: Core Processes and EHR Requirements for Public Health Syndromic Surveillance: International Society for Disease Surveillance (ISDS) 2011 January 11, 2011.