MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Collection and Integration of Clinical Data for Surveillance

William B. Lober[a,b], Atar Baer[c], Bryant T Karras[b,d], Jeffery S Duchin[c]

[a]Biomedical & Health Informatics, School of Medicine, University of Washington, Seattle, Washington USA
[d]Northwest Center for Public Health Practice, University of Washington, Seattle, Washington USA
[c]Public Health – Seattle & King County, Seattle, Washington USA
[d]School of Public Health and Community Medicine, University of Washington, Seattle, Washington USA

Abstract

Objective: The syndromic surveillance project at Public Health – Seattle & King County incorporates several data sources, including emergency department and primary care visit data collected and normalized through an automated mechanism. We describe significant changes made in this "second generation" of our system to improve data quality while complying with privacy and state public health reporting regulations. Methods/Results: The system uses de-identified visit and patient numbers to assure data accuracy, while shielding patient identity. Presently, we have 124,000 basic visit records (used to generate stratified denominators), and 29,000 surveillance records, from four emergency departments and a primary care clinic network. The system is capable of producing syndrome-clustered data sets for analysis. Discussion: We have incorporated data collection techniques such as automated querying, report parsing, and HL7 electronic data interchange. We are expanding the system to include greater population coverage, and developing an understanding how to implement data collections more rapidly at individual hospital sites, as well as how best to prepare the data for analysis.

Keywords:

Software Design, Bioterrorism, Public Health Informatics, Population Surveillance

Introduction

This paper briefly reviews three implementations of Emergency Department (ED) based syndromic surveillance in Seattle & King County, which have been presented previously. These earlier systems were a manual and semi-automated system, as well as a first generation automated system ("Version 1"). The paper then presents the current, second-generation, automated reporting system ("Version 2"), which collects multi-tiered data from ED and primary care visits, and provides automated transmission of normalized data to Public Health for further analysis. The data structures of the Version 2 system are designed to integrate public health surveillance with notifiable condition reporting. This specifications of this system, and the evolution of surveillance systems in our community may be of use to others wishing to implement similar bioterrorism surveillance based on clinical data, and debating the scope and design of such a system

Manual, temporary surveillance

Public Health - Seattle & King County (PHSKC) first implemented Emergency Department (ED) based syndromic surveillance for conditions compatible with bioterrorism and naturally occurring disease outbreaks in the Seattle-King County metropolitan area during the 1999 World Trade Organization (WTO) ministerial, in collaboration with the Centers for Disease Control and Prevention [1]. This surveillance was accomplished through use of a manual, clinician-based, temporary (or "drop-in") system at eight EDs, and collected data on specific symptom complexes ("syndromes"), and on any relationship between the patient and the WTO venue.

The surveillance system operated at all sites from 1 week before to 8 days after the WTO meetings, with over 10,500 ED visits reported. The project established the willingness of hospitals to participate in ED-based public health surveillance and identified areas for improvement in surveillance methods and technology. In particular, we identified development of automated collection and transmission of surveillance data from existing clinical databases as priorities.

Subsequent national events have highlighted the importance of early detection of bioterrorist attacks and disease outbreaks, and a variety of systems have been developed to address this need. This article will review the initial development of an automated ED based surveillance system after the WTO surveillance system ended, and describe our second generation system, which was developed in order to address both privacy regulations and data quality.

Semi-automated system

In Seattle, daily surveillance based on clinical visit began in 2000 at a single large community hospital ED. A daily email transmission of a subset of de-identified patient log data was sent to PHSKC. These visits were identified based on a search of diagnoses and free text from the chief complaint field, against a target list developed by epidemiologists in PHSKC.

Early automated surveillance

The Clinical Informatics Research Group (CIRG) at the University of Washington was engaged to develop fully automated infrastructure for the system and to develop a population-based system by accessing, integrating, and standardizing data from additional EDs in the county.

The Version 1 automated surveillance system, run by CIRG in an application service provider model, was deployed in June 2001 to collect visit level de-identified data from three EDs and a network of primary care clinics [2]. Visits of interest are defined as visits where the chief complaint matches a free text search, or the ICD-9 coded diagnosis matches a list of diagnoses of interest. These lists were developed locally within PHSKC. Data collected on visits of interest includes date and time, age, sex, chief complaint, diagnosis, and disposition. These data elements are consistent with those reported in a survey of similar systems [3]. As of September 15, 2003, we have 183,642 visit records, as detailed in Table 1, which reflects historical data and number of deidentified visit records from each site.

Table 1: Summary of Version 1 Visit Data (as of September 15, 2003)

Institution	Data Since	Visit Records
Hospital 1	May 1992	59213
Hospital 2	January 1997	26420
Hospital 3	March 2000	7973
Primary Care Network	October 1998	90036
Total Version One Records		**183642**

Challenges in data collection

As we used the Version 1 system we became aware of challenges related to missing and duplicate data. Missing data was generally the result of the reporting process having been temporarily disabled on the hospital side. We developed a notification system that sends text pages when data from a particular hospital failed to arrive at the time anticipated. Manual follow-up and requests for retransmission have largely solved the missing data problem.

Duplicate data occurred when the same data were transmitted more than once from the hospital site, particularly when restarting workstations running automated scripts, when re-transmitting missing data, and when testing new interfaces. We could detect and remove duplicate visits on the basis of identifying records with matching data across fields, but at the risk of performing inappropriate deletions in sparse data sets. In addition, we discovered that we could miss patients who were admitted to the emergency department prior to, but discharged after the daily report had occurred. We began to develop an architecture that would allow us to reliably identify and remove duplicate visits, and to collect data covering overlapping time periods, at the same time that hospitals were becoming increasingly concerned about the privacy regulations of the Healthcare Insurance Privacy and Accountability Act (HIPAA)[4]. We decided to address both of these concerns with our new architecture.

Methods

In Washington, in addition to specific notifiable diseases and conditions, health care providers and hospitals must immediately report cases of suspected bioterrorism origin, outbreaks and suspected outbreaks of disease, and cases or clusters of unexplained critical illness or death [5]. The interaction between HIPAA and public health reporting has been described [6]. HIPAA permits reporting to public health without authorization or consent when that reporting is legally mandated. When they contain protected health information, these reports must be disclosed to any patient who requests an accounting from the hospital or provider.

Our strategy is to operate within the framework of legally mandated reporting while collecting only the minimum amount of identifying information necessary. To do this, we defined two de-identified keys, one based on the hospital's visit number, and the other based on the patient number. These keys are generated from a combination of data elements, including the visit number, using a one-way encryption scheme [7]. This ensures that we have a unique way of matching records, but we have no way of deriving the original medical record number or visit number, both of which are considered protected health information under HIPAA.

Three-tier data collection

We defined a new, multi-tier set of data elements to address three different levels of granularity, summarized in Table 2. Tier 1 data is collected for the purposes of creating a stratified denominator, or an accurate count of all visits that may be subdivided by age or sex. These data also include a de-identified visit key to allow us to identify duplicate reports. Tier 2 data is intended for syndromic surveillance, as well as to support the initial phases of outbreak investigation. In addition to de-identified visit and patient keys, these data elements include date of birth, chief complaint and all diagnoses, date and time in and out, disposition, ZIP code, race, attending physician, occupation, and employer. Individual sites may report only a subset of these data. Tier 3 data is not presently being collected, however it is intended to facilitate reporting of individual patients with specific notifiable conditions. These data include the foregoing, with the addition of specific identifiers such as name, medical record number, phone number, and provider contact information.

Table 2: Summary of 3 Tier Data Collection System

Tier One data are used to calculate a denominator for purposes of estimating rates of disease in the hospital/ clinic population, and for stratifying that denominator for subset analysis by age or sex.
Tier Two data are used to monitor population health through syndromic surveillance, and to gather health care utilization information during an outbreak.
Tier Three data are used to supplement existing provider reporting mechanisms for reporting notifiable conditions

Data collection design

We support several architectures for data collection from individual hospitals: automated transmission of text reports and XML documents using encrypted protocols such as HTTPS, secure copy (scp), and Secure File Transfer (SFTP); automated, on-site queries of hospital databases; and transmission and parsing of HL-7 message streams; all of which are depicted in Figure 1. The data are obtained from emergency department systems,

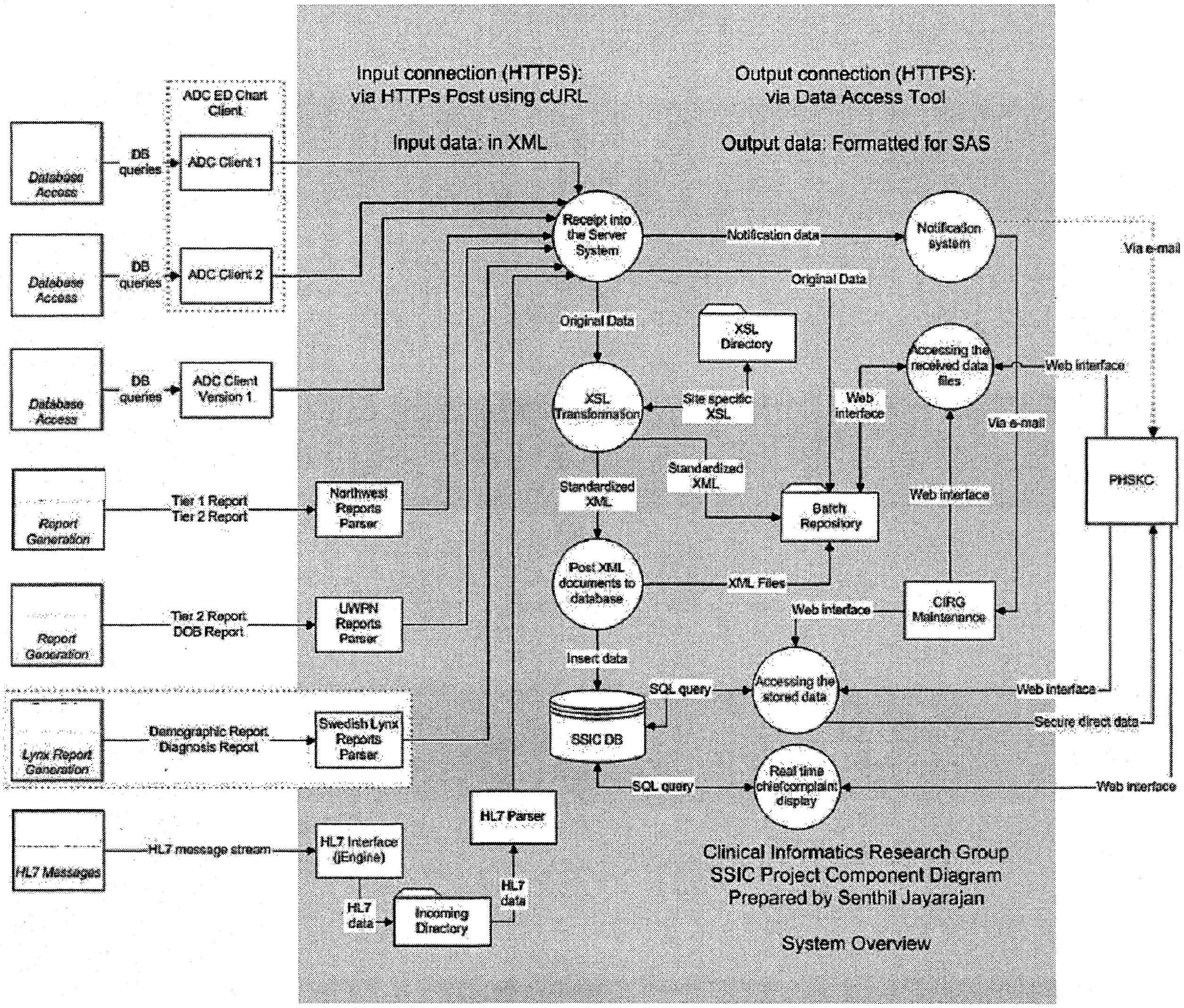

Figure 1 - System Architecture. System data sources appear on the left-hand side of the drawing, and represent individual emergency department or enterprise systems access through direct database queries, through parsing of periodic reports, or though receipt of HL7 messages. Data are processed through custom parsers, or custom data extractors, and encoded in XML documents. On the server, data are processed by a site specific transform for additional semantic transformations, stored in both batch and database format and sent both to a notification system and to Public Health for analysis. This complex system is monitored and tested at several points

enterprise information systems, registration/ADT, and/or billing systems. Once collected, those data are normalized and stored in a centralized database. Automated queries are run on behalf of the PHSKC epidemiologists five days a week. In addition, ad hoc queries may be run at any time for the entire system or for particular hospitals and dates. These queries produce single-table, non-normalized, "raw" data sets, as well as "analysis" data sets (with optional syndrome clustering) designed specifically to work with the Early Aberration Reporting System (EARS) software developed by the Centers for Disease Control and Prevention [8].

Results

The system is implemented as a distributed web application, with web services interfaces between its primary components, as depicted in Figure 1.

This architecture allows these components to be independent of one another and to be easily distributed across a variety of platforms, with clean delineation between the system elements. The central server components are developed on the Linux-based LAMPP platform [9], but presently run on Windows 2000 server as well, using either the Microsoft SQL Server or open-source MySQL databases. We are exploring the development of an open-source version of our software system, for general distribution to the public health community.

Data collection performance

The Version 2 system (incorporating de-identified keys and multi-tier data collection) was deployed beginning in April, 2003. Four hospitals (two that reported previously and two new ones) report through this system, and we are in the final stages of obtaining reliable data from the primary care network. Table 3 describes the data we have collected through September 15, 2003.

Table 3: Summary of Version 2 Visit Data (as of September 15, 2003)

Institution	Data Since	Visit Count Records	Surveillance Records (Tier 2)
Hospital 1	Apr, 2003	107933	60742
Hospital 2	Nov 2001	8559	8559
Hospital 3	Jan 2001	83009	23046
Primary Care Network	Jun 2003	52080	5551
Hospital 5	Mar 2003	43837	2558
Hospital 6	Nov 2002	28205	28205
Hospital 7	Apr 2000	111922	19070
Total Version Two Records		**435545**	**124104**

Table 2 reflects historical data from each site, as well as the number of Tier 1, or visit count records (used to obtain stratified denominators) and Tier 2, or surveillance records (which contain chief complaint, diagnosis, etc., on visits of interest). Our historical record is less complete than in the Version 1 system, however our overall record has grown much larger, with 435,545 Tier one, or denominator reports from the four operational sites, and 124,104 Tier two, or surveillance reports.

Discussion

This project has taught us a great deal about the complexities of collecting clinical data from multiple institutions. We now have experience with a variety of different data collection strategies. These include strategies based on repeated automated queries or report writing and on electronic data interchange messaging. This experience has increased our efficiency in adding new sites, and in handling new data types. We are also developing a more standardized way of approaching different institutions, conveying information about the system, and developing an agreement for the institution to participate in the project. We have found a standardized package of documents, some aimed at policy and others aimed at technical specifications, to be helpful in this regard.

Normalization

Normalization remains a significant challenge. The mechanics of syntactic or structural transformations of data are relatively straightforward. We rely on XML representations of data and on both text parsing and an XML transformation language to help us restructure transmitted data. However, semantic transformations, or transformations of meaning, remain as challenging in public health surveillance as they are in other areas of clinical informatics. We are hindered by the absence of widespread agreement on what data elements to collect and store in clinical systems, and on which coding schemes to use for particular types of data. However, while work practices and coding schemes may be dissimilar, we believe that population based detection may well tolerate these disparities within the data set – a key question we hope to answer during our evaluation of the utility of these data.

Portability

One of our goals is to replicate the same infrastructure across the Linux and Windows platforms. We are now in the process of doing this and believe that early attention to modular design has greatly facilitated this process. At present many of the components individually run on either platform, but we have not yet built a Windows-based complete system. Another goal is the replication of the system outside of the initial local health jurisdiction. We have been working to implement a very similar system in a neighboring three-County region. We now have one hospital online, and plan on adding three others. In doing so, we are learning about variations not just within hospitals, but also within local health jurisdictions.

Standards

Attention to appropriate standards for syndromic surveillance systems is important if such systems are to be widely employed. The CDC has put forth a set of standards for the Public Health Information Network, one component of which is the National Electronic Disease Surveillance System [10]. While our system broadly meets the architectural and security standards, we will move towards adoption of the associated messaging standards as well. Though our implementation is at the level of local public health, we are designing with an eye towards our system participating in a larger state and federal network, with appropriate levels of aggregation and de-identification.

Conclusion

This brief report summarizes the design principles and initial implementation of a pilot automated syndromic surveillance system. Overall the system has performed as expected, meeting the goal of automated collection and transformation of heterogeneous data from multiple institutions into a standardized data set for statistical analysis. Because of the small proportion of EDs currently participating in the system, an evaluation of the public health attributes and practical utility of the system is not yet feasible. Because the ultimate public health value of syndromic surveillance systems has not been established, our current priority is to build a population-based system that can be used to address such questions [11].

Acknowledgements

We would like to gratefully acknowledge the support of the Centers for Disease Control and Prevention through the Washington State bioterrorism preparedness grant (B2 section) U90/CCU017010-02. In addition, we would like to thank Jack Ciliberti, MD for his efforts in initiating information system-based surveillance in King County, Senthil Jayarajan for the preparation of Figure 1, and Jane Sebastian for assistance in preparation of the paper.

References

[1] Plough A. WTO enhanced surveillance project – local and national collaboration leads to success. *EPI-LOG Communicable Disease and Epidemiology News*. Dec. 1999;(39)12

[2] Lober WB, Trigg LJ, Karras BT, Bliss D, Ciliberti J, Stewart L, Duchin JS. Syndromic Surveillance Using Automated Collection of Computerized Discharge Diagnoses. *Journal of Urban Health* 2003 80: 97i-106i.

[3] Lober WB, Karras BT, Wagner MM, Overhage JM, et. al., Roundtable on Bioterrorism Detection: Information Systems-based Surveillance. *Journal Am Med Informatics Assn.* Mar/Apr 2002;9(2):105-115.

[4] National Committee on Vital and Health Statistics. *Uniform Data Standards for Patient Medical Record Information.* Report to Secretary of U.S. Department of Health and Human Services. Health Insurance Portability and Accountability Act (HIPAA) of 1996. Washington, DC: Health and Human Services, 2000.

[5] Washington Administrative Code (WAC 246-101-010)

[6] Broome CV, Horton HH, Tress D, Lucido SH, Koo D. Statutory Basis for Public Health Reporting Beyond Specific Diseases. *Journal of Urban Health* 2003 80: 14i-22i.

[7] Rivest R, The MD5 Message-Digest Algorithm RFC 1321, Internet Engineering Task Force, April 1992. (Retrieved June 25, 2003 from: http://www.ietf.org/rfc/rfc1321.txt)

[8] Hutwagner L, Thompson W, Seeman GM, Treadwell T. The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS). *Journal of Urban Health* 2003 80: 89i-96i.

[9] Dougherty D, LAMP: The open source web platform. 2001. (Retrieved June 20, 2003, from: http://www.onlamp.com/pub/a/onlamp/2001/01/25/lamp.html)

[10]Public Health Information Network web site, Centers for Disease control and Prevention, 2003. (Retrieved June 25, 2003, from http://www.cdc.gov/phin/)

[11]Duchin JS. Epidemiological Response to Syndromic Surveillance Signals. *Journal of Urban Health*: 2003;80 Number 2 (SS1).

Address for correspondence

Bill Lober MD
lober@u.washington.edu
www.cirg.washington.edu