DEVELOPMENT AND EVALUATION OF LEVERAGING BIOMEDICAL INFORMATICS TECHNIQUES TO ENHANCE PUBLIC HEALTH SURVEILLANCE

by

Kailah T. Davis

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

August 2014

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Kailah T. Davis**

has been approved by the following supervisory committee members:

Julio Facelli, Chair **1/27/2014**
Date Approved

Catherine Staes, Member **2/06/2014**
Date Approved

Leslie Lenert, Member **2/03/2014**
Date Approved

Per Gesteland, Member **2/12/2014**
Date Approved

Robert Kessler, Member **6/10/2014**
Date Approved

And by **Wendy Chapman**, Chair of

the Department **Biomedical Informatics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Public health surveillance systems are crucial for the timely detection and response to public health threats. Since the terrorist attacks of September 11, 2001, and the release of anthrax in the following month, there has been a heightened interest in public health surveillance. The years immediately following these attacks were met with increased awareness and funding from the federal government which has significantly strengthened the United States surveillance capabilities; however, despite these improvements, there are substantial challenges faced by today's public health surveillance systems. Problems with the current surveillance systems include: a) lack of leveraging unstructured public health data for surveillance purposes; and b) lack of information integration and the ability to leverage resources, applications or other surveillance efforts due to systems being built on a centralized model. This research addresses these problems by focusing on the development and evaluation of new informatics methods to improve the public health surveillance.

To address the problems above, we first identified a current public surveillance workflow which is affected by the problems described and has the opportunity for enhancement through current informatics techniques. The 122 Mortality Surveillance for Pneumonia and Influenza was chosen as the primary use case for this dissertation work. The second step involved demonstrating the feasibility of using unstructured public health data, in this case death certificates. For this we created and evaluated a pipeline composed of a detection rule and natural language processor, for the coding of death certificates and the identification of pneumonia and influenza cases. The second problem was addressed by presenting the rationale of creating a federated model by leveraging grid technology concepts and tools for the sharing and epidemiological analyses of public health data. As a case study of this approach, a secured virtual organization was created where users are able to access two grid data services, using death certificates from the Utah Department of Health, and two analytical grid services, MetaMap and R. A scientific workflow was created using the published services to replicate the mortality surveillance workflow. To validate these approaches, and provide proofs-of-concepts, a series of real-world scenarios were conducted.

To my loving family.

TABLE OF CONTENTS

Chapters

Appendices

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

CDC	Centers for Disease and Control and Prevention
caBIG	Cancer Biomedical Informatics Grid
DCP	Death Certificates Pipeline
ESSENCE	Electronic Surveillance System for the Early Notification of Community-based Epidemics
GAO	United States Government Accountability Office
HHS	Officials at Human Health Services
ICD	International Classification of Diseases
ILI	Influenza-like Illness
MMWR	Morbidity and Mortality Weekly
MRS	Mortality Reporting System
NBAS	National Biosurveillance Advisory Subcommittee
NBIC	National Biosurveillance Integration Center
NBSHH	National Biosurveillance Strategy for Human Health
NLP	Natural Language Processing
PHI	Public Health Informatics
P-I	Pneumonia and Influenza
RODS	Real-time Outbreak Detection System
UDOH	Utah Department of Health

US	United States
VO	Virtual Organization
gRAVI	Grid Remote Application Virtualization Interface

ACKNOWLEDGMENTS

I would like to thank my committee members for their advice, contribution and encouragement throughout the entire process: Julio Facelli, PhD; Catherine Staes, PhD; Per Gesteland, MD; Robert Kessler, PhD; and Leslie Lenert, MD.

I thank the University of Utah Center of High Performance Computing staff Sean Igo and Ron Price for sharing their expertise and knowledge with me; your assistance has helped me complete different phases of this research project.

I would like to thank my family and close friends for their continuous love, support, encouragement and humor throughout this journey.

This work was generously funded by the grants from the National Library of Medicine training grant T15LM007124 and from the Centers of Disease Control and Prevention Center of Excellence (IP01HK000069-10).

CHAPTER 1

INTRODUCTION

Problem Statement

Public health has been a public health initiative for many centuries [1]. Timely and accurate detection have been stressed because it can dramatically reduce the morbidity and mortality of a population [2]. Moreover, data integration and/or system interoperability is crucial to the early detection of an outbreak or exposure [3]. The rapid spread and economic impact of SARS in 2003 [4] and the global response to the 2009 pandemic influenza (H1N1) outbreak [5] are prime examples of the need for enhancing global public health surveillance systems. Throughout the literature, it is documented that traditional public health surveillance systems in the United States and elsewhere are not effective to fully address these emerging threats and challenges posed by the lack of interoperability and/or system integration in public health [6, 7]. Moreover, current public health systems do not take advantage of recent advances in the field of informatics and computer science, which are commonly used in scientific research, e-commerce, finance, etc. This situation of the disease surveillance systems is unacceptable, especially in today's world where news media and the public expect immediate information on new public health threats. Improving the current system can provide better situational awareness and aid in the execution of an appropriate countermeasure response. Therefore, new nontraditional methods using modern computer science and informatics approaches are critical to address these problems.

Problem Definition

Interest in public health surveillance has been on the rise since the early 2000s terrorist. A result of these attacks was the increased federal government funding, which has significantly strengthened the United States surveillance capabilities; however, despite these improvements, there are substantial opportunities and challenges in sustaining and enhancing our public health surveillance capabilities. For example, today's public health surveillance systems—though extensive and sophisticated in many cases—are unable to fully meet the demands for a comprehensive, near real-time information that is essential for quick, guided responses. A major reason for that may be many of the current technologies used in monitoring diseases are significantly limited in their ability to detect emerging health threats and share data across various channels efficiently among agencies. Strengthening these capabilities are key priorities to the National Biosurveillance Strategy for Human Health (NBSHH) [8]; however, public health officials are facing challenges to enhance surveillance capabilities based on the recommendations of *The Strategy* [9]. Consequently, the Centers for Disease and Control and Prevention (CDC), in collaboration with its national and international partners, has been involved with ongoing research to find new methods and technologies to augment existing public health surveillance capabilities that will help bridge the data exchange gap that currently exists. A 2010 report by the President's Council of Advisors on Science and Technology (PCAST) [10] noted that Health Information Technology lags behind other industries in areas such as data analysis and the integration of data and systems to aid in decision making [10]. New resources and tools are developing in fields such as informatics, computer science and geography which can be leveraged to help in the development of an enhanced public health surveillance system. However, before deploying these technologies in public health departments, it is imperative to demonstrate their usefulness. Gesteland et al. [11], in their groundbreaking EpiCanvas paper, showed that such demonstrations can be done very effectively by developing academic prototypes.

The following three sections describe current technical issues which need to be addressed to enhance public health surveillance.

Problem 1: Unstructured Data

The US public health system is a large, distributed entity that captures a wealth of information. However, much of the data useful for public health surveillance are usually available in unstructured, free text format which can easily be read and understood by humans, but is difficult for computers to decipher [12]. Moreover, the analyses of these unstructured data have inherent problems and should use structured data to successfully identify target population. Structured, coded data such as International Classification of Diseases, 10th Revision (ICD-10) codes are extensively used in by public health surveillance systems because of their ability to easily parse, process, and manipulate data [12]. In addition, it is postulated that the ability to transform these data into a more structured format is ideal for surveillance systems because it provides greater sensitivity over specificity allowing greater detection of events of interest [12].

Potential data sources for public health surveillance can be easily found unstructured, free-form text throughout public health organizations [12] such as vital records systems or epidemiology investigator's case notes. These sources add value to a public health surveillance investigative team because they provide information that are not typically available in structured data sources [12, 13]. Therefore, the ability to structure this information for surveillance purposes will allow for greater sensitivity provided by structured data and increase specificity in detecting an outbreak.

Most public surveillance systems that use unstructured, free text data are usually from information from emergency departments, and/or clinical reports such as discharge summaries, radiology reports, the patients' electronic medical record, and more [13]. These textual sources provide potentially valuable data to aid with the detection and characterization of diseases and/or emerging public health threats. Nevertheless, despite their success, the use of unstructured data is not prevalent among public health systems [8] primarily because of the limitation in the nontraditional data sources many of these systems collect [14].

One source of data which has proven effective and is widely used in public health is death certificates. The causes of death from death certificates are used weekly to determine the number of deaths that are associated with pneumonia and influenza[12]. This information is reported by the CDC (http://www.cdc.gov/flu/weekly/) to assess the severity of the pneumonia and influenza season in the United States. Persons have alluded to the idea of mining information from death certificates for conditions of interest [12]; however, current research does not provide information on processing and coding death notes to help create a more suitable, structured, coded death data for public health surveillance systems to decipher [15]. Therefore, rigorous research is needed to develop the capability, and evaluate the feasibility of utilizing unstructured public health data to contribute to the earlier detection of conditions and threats.

Problem 2: Integration of Information

Most public health departments are "organized vertically" into data silos and as a result, public health information exists in many public health systems: immunization registries, birth and death information systems, individual case reports, data files and surveillance systems [16, 17]. Many of these were developed to meet the information needs and goals of individual programs [16, 17]. Although their implementations have been successful in allowing diverse programs to collect the necessary information the lack of integration has fostered a "patchwork of data systems" that has resulted in duplication of efforts, unutilized public health information, and difficulty in efficiently and effectively performing the three primary functions of public health: assurance, policy development and assessment.

Transforming public health's disease surveillance system to develop a more robust, integrated system (to share information among agencies) has become increasingly important [3]. Multiple laws and initiatives such as The National Biosurveillance Integration Center (NBIC) code [18], the Federal Health Architecture initiative [19], HPSD-21 (Public Health and Medical Preparedness) [20], Homeland Security Presidential Directive (HSPD)-10 (Biodefense for the 21st Century) [21], Public Law 109-417 (Pandemic All-Hazards and Preparedness Act) [22], Public Law 110-53 (Implementing Recommendations of the 9/11 Commission Act of 2007) [23], and most recently, the National Biosurveillance Strategy for Human Health [8], have all called for the protection of the health of Americans by facilitating information sharing across all sectors and jurisdictions in regard to health-related threats and events.

In response to these legislative initiatives and recommendations from various committees, the federal government has invested millions of dollars in the development and implementation of numerous separate and independent disease surveillance and biosurveillance systems. Despite the millions invested, there is still lack of integration and interoperability among these systems. This was pointed out by The National Biosurveillance Advisory Subcommittee (NBAS) which was created to review federal funded biosurveillance efforts [24]. In their 2009 review, the NBAS recorded a total of 300 disparate biosurveillance efforts and concluded that many of these systems were program/disease specific, duplicative and lacks integration and/or interoperability. Although many initiatives came about after 9/11 to increase the sharing of public health information since the inception of this dissertation very little progress has been made. Academic proof-of-concept prototypes are an alternative approach to engage the public health community with modern approaches such as web services and grid technology which are widely used in the research academic setting.

Problem 3: Sharing of Resources and Applications

As noted earlier, the federal government has invested in developing and implementing equipment and technologies to support early detection of diseases. However, during that period there was, and still is, no comprehensive national strategy to provide guidance on how to integrate analysis mechanisms, resources and technologies developed by different federal agencies [25].

Currently, public health functions in the United States are conducted at 59 state and territorial health departments, over 3,000 local health departments and other entities such as clinical laboratories and federal agencies [26]. Despite the many participants involved, there is little to no communication and collaboration across the multiple public health agencies and government sectors. Officials at Human Health Services (HHS) and other federal, state, and local agencies recognizes the need to increase the use of information technology in public health systems to collect, analyze, and share public resources [26]. However, it is noted that the HHS has not developed or implemented an overall strategy to integrate disparate public health systems, thus not achieving the unified, nationwide, electronic public health network that was a stated in the Pandemic and All-Hazards Preparedness Act [22]. Moreover, the Government Accountability Office (GAO) has noted, public health agencies are limited in their ability to electronically collect, analyze, and, share public health information [26]. Therefore, it is evident that providing an infrastructure which will enhance and foster collaboration among federal, state, and local government agencies can reduce the duplication of efforts and provide resources to help share not only data but also analytical capacity.

Study Justification

Fiscal, cultural, social and political issues in public health hamper the enhancement and/or development of surveillance system infrastructures to address emerging threats and challenges in a timely manner. Such a system will allow public health officials dynamic access to tools and disparate, heterogeneous information to meet their ever-changing needs. However, over the past decade, the literature has shown that it is difficult to create this type of public health surveillance system [27]. Nevertheless, there are opportunities to leverage current methods and technologies to overcome current challenges faced in enhancing and transforming our public health surveillance systems capabilities. Therefore, the dissertation aims to demonstrate how different technology can be leveraged to overcome three common challenges faced by current public health surveillance systems. In addition, questions posed by the NBAS report about leveraging resources from other fields to build such a system can be answered. Table 1.1 presents a summary of the key challenges that are investigated in this dissertation and what methods can be leveraged from the informatics domain to address these challenges are approached. A background of each technology/method is found in Chapter 3.

Table 1.1: Description of key problems and corresponding action being taken

Key Problem	Method being leveraged	Approach Taken
Public health unstructured data	Natural Language Processing (NLP)	Use a natural processor to generically code death certificates and apply data mining methods to identify cases of interest (pneumonia and influenza case).
Information sharing	Grid technology	Explore and develop a decentralized architecture for sharing death certificate records.
Application and resource sharing to increase collaboration and decrease duplications.	Grid technology and R	Use grid technology to allow users access to analytical services that are applicable to the public health domain. These applications allow access to NLP tools in a non-technical environment and data analysis using an open source statistical software package.

Main Objectives

This project aims at addressing key challenges that limit the enhancement and transformation of public health surveillance systems by providing a glimpse of what can be accomplished at national scale when using advanced computational methods to provide a nationwide public health network for public health surveillance. This research study aims to answer the following questions:

Primary question. How can we leverage existing technologies to develop prototypes as proof-of-concepts so that these technologies can be used to enhance current surveillance systems?

Secondary question. What existing technology can be used to:

1. Create a pipeline to process unstructured public health data and automatically identify conditions/diseases of interest.
2. Create a robust, interoperable information-sharing between public health agencies.

Study Aims

Aim I

Demonstrate the capacity to use death certificates as an example of using unstructured public health data for surveillance using available off-the-shelf technologies.

Hypothesis. A death certificates pipeline (DCP), composed of a detection rule and a natural language processor, can be created for the 1) standardization/coding of death certificates, and 2) automatic identification of pneumonia and influenza cases. It is hypothesized that the DCP accuracy will be comparable to existing methods.

Research question 1.1. Can natural language processing tools be used to standardize and code death certificates to identify pneumonia and influenza deaths?

Research question 1.2. How does the output compares to current standard methods?

Aim II

Evaluate the use of grid technology to share, analyze and process public health data.

Hypothesis. A new public health surveillance architecture can be developed to allow the sharing of public health data and provide access to distribute computing tools not commonly used in public health to aid in the processing of public health data.

Research question 2.1. Can grid technology be used to effectively share electronic death records?

Research question 2.2. Can grid technology approaches be leveraged to aid in processing public health data, specifically death certificates, in a less technical environment?

Aim III

To develop and deploy commonly used public health surveillance algorithms in a grid environment.

Hypothesis. Exposing an analytical tool based on the statistical open source tool R can be developed to demonstrate the usefulness of grid technology to share resources that were developed by different public health entities.

Related Work

This section discusses the different approaches which have been used to help overcome the key challenges being addressed in this dissertation.

Unstructured Data

The importance of unstructured or textual data for monitoring and identifying various conditions and diseases—especially globally in the private sector—is growing. For example, sources like the Public Health Agency of Canada's Global Public Health Intelligence Network (GPHIN) [28] constantly mines available free text global news media sources and health and sciences websites for disease outbreaks. Similarly, Project Argus [29] and ProMed-mail (the Program for Monitoring Emerging Diseases) [30] automatically monitor local and national media reports to rapidly disseminate information on outbreaks. HealthMap [31] and Global Health Monitor [32] also use internet and media information such news reports to automatically provide users potential disease cases about human and animal health.

The United States federal government has also built systems that incorporate unstructured data into their analyses many of which have used chief complaints; for example, the Biosense system uses natural language processing techniques to extract data from unstructured chief complaints [33]. Another system, the Biosurveillance Common Operating Network (formerly the National Biosurveillance Integration System (NBIS)) integrates information from over 120 data streams and handles a large quantity of structured and unstructured data including open source news feed.

While the approaches discussed are novel, and idealistic, many of the attempts to use unstructured data in public health surveillance systems have focused on few data sources, mostly, medical records—chief complaints and nontraditional data sources—news feed. The desired functionality of unstructured data in public health information systems needs to be from various data sources such as data from public health agencies. As the NBSHH noted, more evaluation is needed to demonstrate the usefulness (particularly the precision supported by the data) of using various unstructured public health data for disease surveillance purposes [8].

Information and Resource Sharing

The lack of information integration presents major vulnerabilities in the public health system's ability to rapidly detect public health emergencies. In 2003, the CDC led the initiative to create a system called BioSense with its aim to mitigate this existing vulnerability [34]. BioSense collects and transmits patient related data such as chief complaints and diagnosis data to a centralized data warehouse. Users can then access data that are being shared to perform various statistical analyses. Despite the attempt to increase information sharing, this centralized model of information sharing has proven to be difficult to scale and costly to maintain [35]. These barriers have been continuously reported in the literature; however, it does not offer a strategy to address those barriers to create an information-sharing environment. The enhancement to the current functionality in existing biosurveillance applications (e.g., BioSense) with distributed computing, grid architectures, and/or an open federated model for data access and exchange is one possible approach that could be used to potentially decrease many of the current limitations.

The Centers for Disease and Control and Prevention (CDC) and their national and international partners have been exploring new methods to increase information sharing in public health. They made poison data available at a national scale by creating a secure data web service; this service used by an in house tool, Quicksilver, to visualize aggregated clinical effect data both geospatially and via a time series chart [35]. Future work and recommendations for the project discussed that the next steps would involve extending the existing web services into grid enabled services. However, to date further work on this project has not been published.

One of the key challenges the literature does not address is the sharing of public health resources. To date, there has not been any research on sharing public health applications for the larger public health community.

Division of Chapters

This introductory chapter, Chapter 1, provided an overview of the core research problem that is being addressed by this dissertation and provides an overview of previous approaches that have been undertaken for this task. In addition, the goals of the study and the hypothesis for each of the study's three aims to help improve upon previous approaches were also presented. To help guide readers through the rest of this dissertation the list below provides an overview of each of the remaining chapters:

- Chapter 2, The Literature Review section describes the public health surveillance needs that motivate this study, prior research that has been conducted in this area, and how this study aims to improve on prior work. This section puts the research project in context and shows the value added by this dissertation.
- Chapter 3, the Technology Background section, describes the technologies used in this research project.

- Chapter 4, Methods section, describes the data and resources used for this dissertation and outline the study's experimental design; additionally, this section describes the informatics procedures that were used to develop the prototypes and details at a technical level and their advantages over other approaches.
- Chapter 5, Results section, contains the findings of this study in substantial detail.
- Chapter 6, Discussion section, provides additional interpretation of the results, explains limitations of this study and suggests opportunities for future work.
- Chapter 7, the Conclusion, concludes the study and postulates on the potential implications of this study for the biomedical informatics research community.

CHAPTER 2

PUBLIC HEALTH SURVEILLANCE SYSTEMS AND INFORMATICS' IMPACT ON PUBLIC HEALTH SURVEILLANCE

Introduction

With public health disease surveillance dating back to the 1348 bubonic plague epidemic [1], there is abundant literature on biosurveillance and public health surveillance. Therefore, for the purpose of this literature review, we will focus on the following:

1. Brief history of Public Health Surveillance and Informatics Influence
2. Literature discussing the integration and coordination of current surveillance systems
3. Literature pertaining to existing surveillance and biosurveillance systems

Brief History

While epidemiology can trace its origins to Hippocrates' study of the relationships between environmental factors and disease, the "Black Death," which was a phrase coined due to the 1348 bubonic plague epidemic, is recognized as one the first public health actions—quarantine—attributed to public health surveillance. The Venetian

Republic appointed three “guardians of public health” to prevent ships with infected passengers from docking at the port [1]. Quarantine measures were used again in 1377 in Marseilles to control the spread of infectious diseases by detaining travelers from plague infected areas [36]. Since then the concept of public health surveillance has evolved over the centuries.

The 15th-century Renaissance brought about the scientific revolution and in 1532, the town council in London, England birthed the idea of systematic, ongoing collection of mortality data, later known as Bills of Mortality, to record the number of persons dying from the plague outbreak [37]. Although these Bills of Mortality data were collected intermittently for over 100 years, it was not until the 1600s that they were used for surveillance purposes when the number of causes of death and burials were reported by the clerks of London to the Hall of the Parish Clerk’s Company. The statistics from the data were published in a weekly Bill of Mortality [38].

In the late 1600s John Graunt comprehensively analyzed and interpreted these weekly bills and in 1662 he published his book Natural and Political Observations Made upon the Bills of Mortality [39]. His revolutionary work allowed him to be cited as the first to “conceptualize and quantify the patterns of disease and to understand that numerical data on the population could be used to study the cause of disease” [1].

In the United States, the first documented legislation for public health surveillance came in 1741 when the colony of Rhode Island passed an act which required the reporting of contagious diseases by tavern keepers. Two years later, a broader law was passed which included the mandatory reporting of cholera, yellow fever, and smallpox [40]. This started the idea of infectious disease reporting being mandated by legislation or state laws.

A few years after the legislation passed in Rhode Island, public health surveillance began being part of policy development. In 1766, Johann Peter Frank encouraged a more extensive form of public health surveillance in his native country Germany, particularly in the areas of schoolchildren health, maternal and child health, injury prevention and public water and sewage disposal [40]. He created a health policy that impacted the people of Germany and countries which had close cultural contact with the country [1]. Many countries, such as France in 1788, soon recognized that the health of the population was the state's responsibility [1].

During the 19th century public health surveillance came into its own due to pioneering work done by Sir Edwin Chadwick, William Farr, John Snow, and Lemuel Shattuck. As the Secretary of the Poor Law Commission in England, Sir Edwin Chadwick used surveillance data to show that there was a correlation between poverty and disease, using surveillance data [41]. Similar findings were reported by Lemuel Shattuck in his "Report of the Massachusetts Sanitary Commission" (1850), where he used survey data from sanitary conditions in Massachusetts to show the link between communicable diseases and living conditions and maternal and infant mortality [41]. In his report, Shattuck recommended that a permanent statewide public health infrastructure should be created and suggested the establishment of health offices at state and local levels to assist in the gathering of statistical information on public health conditions. Shattuck also recommended a statewide decennial census, standardizing causes of death nomenclature, and the collection of health data by gender, age, profession, socioeconomic level and locality [41]. Although his comprehensive plan was not initially adopted, some of his proposed solutions were implemented and became routine public health activities during the 20th century and remain the necessary pillars for implementing any modern public health surveillance system.

In 1838, William Farr was appointed as the first Compiler of Abstract (medical statistician) and during his time at the General Register Office, which was established to improve the accuracy and completeness of mortality data that were collected in England and Wales, he created a surveillance system which led to his being recognized as the "founder of the modern concept of surveillance" [42]. He was the first Compiler of Abstract to collect and analyse vital statistics data (such as birth and death information) to describe the magnitude and impact of a disease in various populations; this evaluation was reported regularly to English authorities and the general public [42].

While William Farr is known as the "founder of the modern concept of surveillance," another British contemporary, John Snow, is known as the father of modern epidemiology for his work in 1854 about the causes of the 19th-century cholera epidemics. Snow plotted the cholera deaths and was able to trace the deadly outbreak in London to a contaminated water pump on Broad Street [38]. Snow described his findings as:

> On proceeding to the spot, I found that nearly all the deaths had taken place within a short distance of the [Broad Street] pump. There were only ten deaths in houses situated decidedly nearer to another street-pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pumps which were nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street…. With regard to the deaths occurring in the locality belonging to the pump, there were 61 instances in which I was informed that the deceased persons used to drink the pump water from Broad Street, either constantly or occasionally…. The result of the inquiry, then, is, that there has been no particular outbreak or prevalence of cholera in this part of London except among the persons who were in the habit of drinking the water of the above-

> mentioned pump well. I had an interview with the Board of Guardians of St James's parish, on the evening of Thursday, 7th September, and represented the above circumstances to them. In consequence of what I said, the handle of the pump was removed on the following day [43].

The pump handle was removed by Snow on September 8, 1854 and this resulted in the rapid decrease of cholera cases. Snow's work illustrated how to use public health surveillance activities for targeting and monitoring public health interventions.

By the 20th century public health surveillance systems expanded and became more diverse and sophisticated. During this time, The United States (US) played an integral part in the development of new models and concepts for public health surveillance. This was due to the fact that the country in 1942 created the Centers for Disease Control and Prevention (CDC) which was responsible for the nation's public health surveillance efforts. By 1955, a little over a decade since its establishment, the CDC intensified its active surveillance of acute poliomyelitis cases to provide evidence that the epidemic could be traced to one vaccine manufacturer [38].

The late 1900s saw many victories for public health surveillance efforts; it played a vital role during the campaigns for smallpox (1967-1988), poliomyelitis and guinea worm eradications and for the control of emerging and reemerging infectious diseases [38]. Also, in the 1980s the advent of microcomputers revolutionized the way many surveillance organizations, such as CDC, collected, analysed, and shared public health surveillance data.

Mortality Surveillance

Much of the investigation uses mortality surveillance as the primary use case. This section presents a brief background of mortality surveillance and the rationale for it being chosen as the primary case study.

As seen in the previous section, the ongoing monitoring of mortality is not a novel idea. For many years it has played a crucial role in detecting and estimating the magnitude of a disease during epidemics or the severity of an influenza season [1]. Today, mortality surveillance continues to be a critical activity for public health agencies throughout the world [44-46].

The US uses death certificates as the primary data source to exemplify epidemics and measure the severity of influenza seasons [47]. Although there are three systems that monitor influenza-related mortality, one system in particular will be the focus of this study, the 122 Cities Mortality Reporting System [46]. This system calculates and publishes the total number of death certificates filed in 122 US cities, as well as providing the number of deaths due to pneumonia and influenza [46]. The current workflow for identifying these deaths includes three key stages: access to death information, identifying pneumonia and influenza deaths from unstructured death records and the use of SAS or manual review of reports to analyze trends; the workflow process is depicted in Figure 2.1.

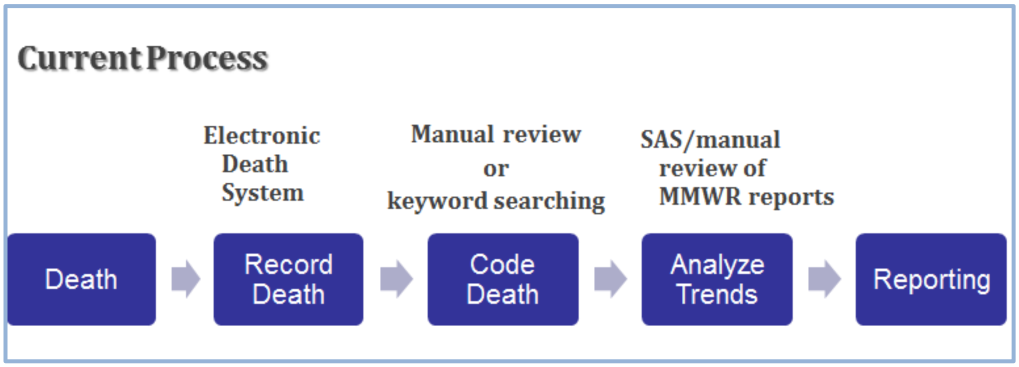


Figure 2.1: Current mortality surveillance workflow

Public Health Informatics Influence on Public Health Surveillance Systems

The CDC's acquirement of its first mainframe computer has led to the increased use of information technology for public health surveillance purposes. Currently, over 500 million dollars per year is invested by the CDC in information systems for a wide range of public health functions [48]. As a result, the CDC is seen as one of the world's leading health information technology (HIT) adopters and a pioneer in the field of public health informatics [48]. Perhaps what contributed to such increase use of information technology is the emergence of public health informatics (PHI). Public health informatics is cited to be modeled after prior "informatics" fields such as medical informatics (the intersection of computer science, information technology and healthcare) and bioinformatics (intersection of computer science, mathematics, biology and statistics). Similarly, PHI is an interdisciplinary field which merges a variety of basic and applied sciences field such as information science, computer science, statistics and public health to improve public health practices, research, learning and the outcome of the population [49]. Using informatics strategies and standards, PHI tools are developed and deployed across public health organizations to provide public health practitioners access to information and tools in timely and secure manner to help guide public health action. Therefore, this section discusses key initiatives led by the CDC to advance surveillance capabilities through the use of information technology.

The emergence of PHI added value and sophistication to modern day public health surveillance systems by improving the timeliness of detecting health threats and providing a more complete and efficient means of health communications; this can be seen through many initiatives by the CDC in the early 1990s to today (See Figure 2.2 for a timeline of key milestones). For instance, early PHI initiatives at CDC were focused on the collection, analysis and dissemination of data. An example of a product of these initiatives is the Epi Info tool which is a statistical program that allows rapid questionnaire development, data entry, and analysis during outbreak investigations [50, 51]. By the early to mid-1990s with the growing popularity of the internet, the CDC's public health informatics initiatives were primarily focused on linking users and developing an integrated public health environment. As a result, CDC PHI initiatives were focused on developing interoperable, standards-based, enterprise architectures.

In 1990, the CDC released two pioneering PHI tools which have revolutionized the way public health practitioners share and access data. The first was the National Electronic Telecommunications System for Surveillance (NETSS) which evolved from a previous pilot project, the Enhanced Surveillance Project [52]. NETSS was one of the first PHI tools used to by state health departments to electronically report cases of notifiable diseases to CDC over a dial-up modem network. Leveraging new technologies, this PHI tool, NETSS (which supports the National Notifiable Diseases Surveillance System), greatly improved the accuracy of the data and timeliness of public health surveillance. During that year CDC also released the Wide-ranging OnLine Data for Epidemiologic Research (WONDER) [53]. This innovative tool allowed CDC epidemiologists to analyze an array of secondary public health data by having access to data on a CDC mainframe.

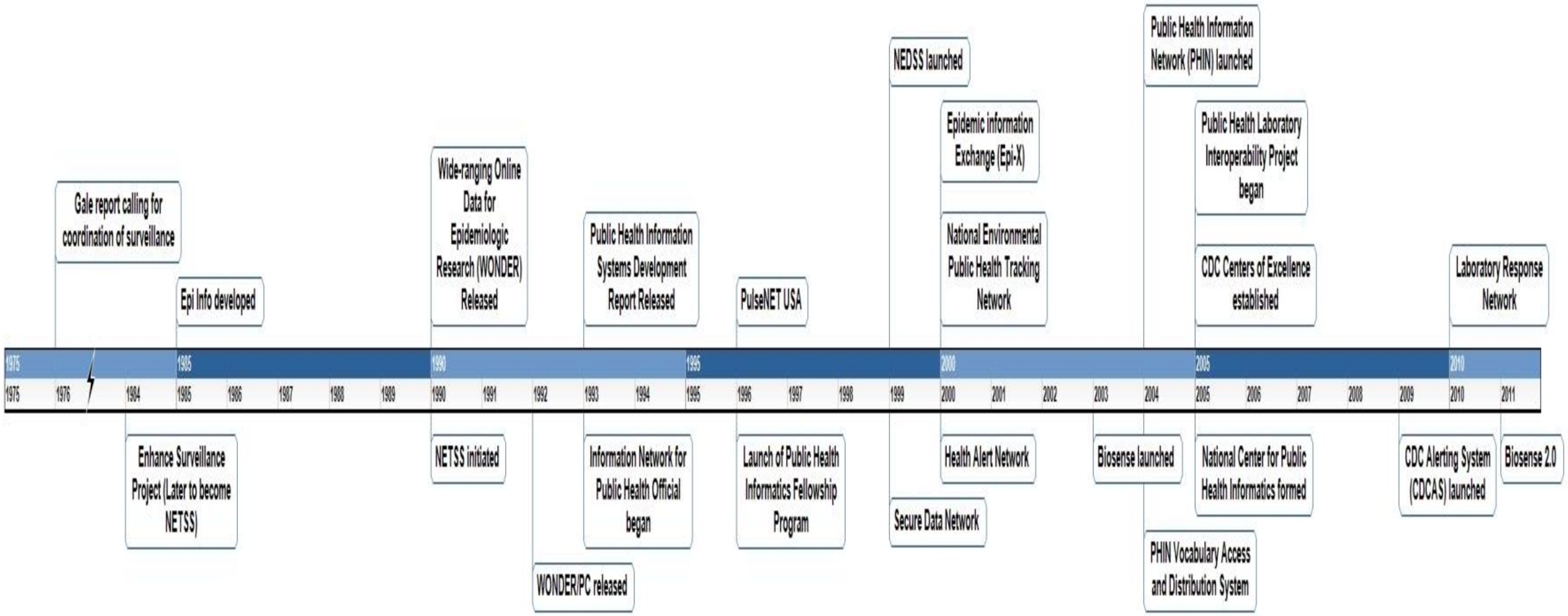

1975
1985
1990
1995
2000
2005
2010
1975
1976
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Gale report calling for coordination of surveillance
Enhance Surveillance Project (Later to become NETSS)
Epi Info developed
Wide-ranging Online Data for Epidemiologic Research (WONDER) Released
NETSS initiated
WONDER/PC released
Public Health Information Systems Development Report Released
Information Network for Public Health Official began
PulseNET USA
Launch of Public Health Informatics Fellowship Program
NEDSS launched
Secure Data Network
Epidemic information Exchange (Epi-X)
National Environmental Public Health Tracking Network
Health Alert Network
Biosense launched
Public Health Information Network (PHIN) launched
PHIN Vocabulary Access and Distribution System
Public Health Laboratory Interoperability Project began
CDC Centers of Excellence established
National Center for Public Health Informatics formed
CDC Alerting System (CDCAS) launched
Laboratory Response Network
Biosense 2.0

Figure 2.2: Timeline of key milestones for informatics at CDC

Just two years after the first release of WONDER in 1992 CDC released WONDER/PC to the broader public health community. This version had enhanced functionalities such as private bulletin boards, an email engine, interactive graphics, and the ability to download confidential data [53]. The WONDER/PC tool revolutionized the way public health data were acquired and analyzed.

Through findings and lessons learnt from early initiatives, and by assessing the then current systems, the CDC and its partners recognized the costliness of surveillance systems particularly for health conditions and categorical diseases. During this time the Internet was gaining popularity; as a result, the CDC assessed the capability of the Internet to provide more efficient approaches to public health surveillance data acquisition, collection and analyses. This resulted in the 1993 initiative to build and develop an integrated surveillance architecture [54]. The implications of these initiatives changed the public health surveillance system architecture because it moved away from stand-alone, centralized solutions for specific problems to networked, integrated solutions that were interoperable with focus on standards-based data exchange [55, 56]. Six CDC PHI initiatives reflect this vision and demonstrate breadth of informatics opportunities for transforming public health surveillance process cycle (data collection/acquisition, data analysis and interpretation, and information products/dissemination) to provide a real-time data flow to support public health activities such as decision support and collaboration: PulseNet USA (14); the National Electronic Disease Surveillance System (NEDSS) [57]; the Epidemic Information Exchange (Epi-X); the Health Alert Network (HAN) [58, 59]; BioSense [34]; and the Public Health Information Network (PHIN) [60]. Table 2.1 provides a broad description of these systems.

Table 2.1: Key CDC PHI initiatives and functional areas

System	Year	Description	Informatics Opportunity
PulseNet USA	1993	The molecular surveillance network for early detection of foodborne disease outbreaks	Information Distribution and Interactive Communication
NEDDS	1999	Standards-based exchange approach for the data exchange of notifiable diseases from health departments to CDC.	Information Distribution
Epi-X	2000	CDC secure internet portal for practitioners to communicate and share information.	Interactive Communication
Health Alert Network	2011	CDC's primary method to share urgent information with the public health network, such as clinicians, public health practitioners and public health laboratories.	Interactive Communication
Biosense	2003 Relaunched in 2011	Improve the nation's capabilities to share information among local, state and federal public health agencies for real-time biosurveillance and situational awareness.	Data Collection, Analysis and Warehousing
Public Health Information Network (PHIN)	2004	A national initiative aimed at improving interoperability among public health information systems	Business Operations

It is important to note that although the CDC has many successes improving public health practice through public health informatics innovations, a recent survey conducted by the organization (434 respondents from the Surveillance Science Advisory Group (SurvSAG)) pointed out that only 22% of respondents are confident in CDC's surveillance systems to perform well in today's world of information technology. In addition, only one in five believed that CDC's surveillance systems are flexible and can readily adopt new information science methods in a rapidly changing environment. Despite dissatisfaction with current CDC surveillance systems many of the respondents (60%) do believe the agency can improve surveillance systems by providing and supporting a standard informatics surveillance system framework across agencies [61]. Moreover, respondents were dissatisfied with CDC's lack of timely access to internal surveillance data through centralized data basses (20%) and public uses datasets from surveillance activities (25%) [61].

From the low agreement rate, one can speculate that it is because many of the systems do not take advantage of technologies being commonly used in different fields. Therefore, future systems will need to include innovative solutions to advance public health surveillance systems.

Existing Public Health Surveillance Systems

While the previous section discussed systems which were developed through the CDC's public health informatics initiatives, there are three major systems in the United States, two of which were not developed by the CDC, which should be discussed in further detail. It is important to mention that while other countries have increasingly been developing biosurveillance systems with substantial capabilities and effectiveness, we will only be focusing on US systems because this dissertation addresses problems faced by the current US infrastructure, and also it would allow a salient overview.

- Real-time Outbreak and Disease Surveillance (RODS): The RODS system, which was developed by the University of Pittsburgh in 1999, is a biosurveillance system that collects real-time surveillance data for early outbreak detection [62]. RODS is connected to hundreds of hospitals' emergency departments, both nationwide and internationally, primarily for syndromic surveillance purposes. RODS is one of the first biosurveillance systems to use innovative data streams for surveillance purposes; the system collects chief complaints, admission records, and over-the-counter drug sales data in real-time. Although in recent years, research on the algorithms used by RODS has tapered off, the project's open source nature increases the possibility of continuing support and development [63].
- Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE): In 1999, the Department of Defense (DoD), in collaboration with Johns Hopkins University Applied Physics Laboratory, developed ESSENCE and installed the system at over 300 military treatment facilities worldwide for the daily monitoring of infectious disease outbreaks. The system provides a visualization and analysis of hospital visits, over-the-counter drug sales and physician office visits data to identify potential outbreaks [64]. Although ESSENCE is primarily used by the DoD, ESSENCE is currently being used by some health departments in the US.
- BioSense: BioSense was developed and operated by the Centers for Disease

Control and Prevention with the intention of creating the first nationwide public health syndromic surveillance system. Initially, BioSense collected and monitored diagnoses and procedure data from the DoD and Department of Veterans Affairs, as well as LabCorp lab test results [7, 34]. Using the collected data, like ESSENCE, it provided some statistical analysis and visualization capabilities for outbreak detection and situational awareness. BioSense's primary objective was to "expedite event recognition and response coordination among federal, state, and local public health and healthcare organizations" [65]. Although BioSense had some functionalities to support national collaboration and comparison across jurisdictions, in 2006, the CDC recognized that BioSense was not successful and began an analysis of the system in hopes of identifying key areas for improvement. Through this analysis, it was identified that many public health practitioners used the system for data exploration rather than for the purpose of detecting outbreaks, due to the system's inflexibility and other limitations [66]. Currently, BioSense is being redesigned through the collaboration of private organizations to develop distributed, cloud-based version of the system [67].

Other Systems and Systems Proposals

While the three systems described above (BioSense, RODS, and ESSENCE) are the largest and most significant, many other organizations and city and state public health departments have developed their own systems. To discuss the exhaustive list of available disease surveillance systems would be tedious and overwhelming; therefore, please refer to the 2010 GAO Report, *Biosurveillance: Efforts to Develop a National Biosurveillance Capability Need a National Strategy and a Designated Leader* [25] for one of the most extensive review of federal level biosurveillance and public health surveillance systems, and for a comprehensive list of national and international disease surveillance systems, The Stimson Center report, *New Information and Intelligence Needs in the 21st Century Threat Environment* [68]. However, one system deserves to be discussed because it uses second-generation Internet systems based more on search engines and social networks. Such techniques are nonexistent in many current public health surveillance systems.

The recent popularity of the Internet has allowed for a new era of surveillance systems. Some are based on localized search engines, such as Yahoo, Bing, or Google, or tracking Social Media content such as that on Facebook or Twitter. These systems capitalize on the wide-spread use of the Internet by the general population. Billions of search terms are being collected and analyzed on a daily basis. Google, Inc. was one of the first entities to develop such a system, namely Google Flu Trends. They were able to isolate the top 45 search terms for influenza-like-illness (ILI). These terms were found to be the most highly predictive (among fifty million terms entered over the lifetime of the search engine) of the ILI counts as reported by the CDC [69].

As noted in the previous paragraph, Google Flu Trends analyzes large numbers of user search queries to track ILI in the population. This method is based on the notion that "there is a correlation between the frequency of certain queries and the percentage of physician visits in which a patient presents with influenza-like symptoms" [69]. The Google Flu Trends algorithm used for ILI prediction is a regression-based model [69], and through a validation experiment using CDC reports, Google Flu Trends predictions were found to be closely related to the CDC's ILI cases with a correlation coefficient between 0.95 and 0.97. It was also found that Google Flu Trends consistently predicted the ILI percentages 7 to 14 days ahead of the CDC's Morbidity and Mortality Weekly Report (MMWR). The validation process of the Google's system also found that when Google Flu Trends data are compared with official influenza surveillance data from other countries, the estimates for search queries about "flu" were very similar to estimates that used traditional flu activity indicators. Google Flu Trends evaluation for different diseases, such as West Nile Virus, Avian Influenza, and Respiratory syncytial virus, all showed good correlation between search volume and official data; however, the correlations were weaker than for influenza-like illness.

Although Google Flu Trends provides valuable information in a quick manner, there are areas where this type of surveillance is limited. Google Flu Trends depends heavily on Internet search terms; as a result, it would be difficult to apply this model to less "popular" diseases as there might not be sufficient search terms to train the model to produce a well-correlated graph. On the other hand, Flu Trends and other similar systems based on social networks and trends are susceptible to "noise" from media coverage; as a result, diseases that have high media coverage for a particular time will result in the increase of "counts" for that disease, not reflecting the actual number of cases. Moreover, Google Flu Trends requires a large population of Web search users, which means that Flu Trends would not be suited in underdeveloped countries or for regions sparsely populated or when the use of the Internet is highly stratified by either sociological or demographic factors. Therefore, the findings would not be generalizable.

Other limitations of the Google Flu Trends system became evident during the 2009 H1N1 pandemic. While the system tracks "in season" epidemics very well, particularly for influenza epidemics, it has limitations when tracking diseases that are not "in season" and lacks in its ability to handle search behavior. It is cited in the literature that Internet search behavior changed during H1N1, particularly in the categories "influenza complications" and "term for influenza." Changes in health-seeking behavior, coupled with the fact H1N1 began in the summer rather than winter, have played a part in the increased influenza-like illness spike in the Google Flu Trends system, causing the correlation between Google Flu Trends and the CDC to decrease substantially. In a recent interview, Lynnette Brammer, a flu epidemiologist with the CDC, spoke about the role of technology in disease tracking and commented that Google Flu trend is a useful tool and complementary to the CDC's influenza surveillance system, but it would not replace the CDC's system in the near future. However, she did mention that real-time information is one advantage the system has over many current systems.

Limitations of Current Systems

While the systems above have many success stories, particularly during high-profile events, their ability to provide reliable and timely early warning is yet to be established. Moreover, although these surveillance systems provide information flows that did not exist before [70], from the literature, it is evident that these systems lack capabilities that users would like to have in current systems such as data-sharing capabilities or access to applications or algorithms that can improve their current processes. The Biosense redesign is being developed to mitigate some of these challenges and recent efforts in 2013 are being investigated to move legacy biosurveillance application models such as ESSENCE [71] from centralized and historically siloed environments to federated and distributed models; however, Patel et al. noted that further evaluation of such resources are needed [71]. Furthermore, the 2013 National

Biosurveillance Science and Technology Roadmap noted that while progress is being made, continuing progress will require focusing investments and coordinating efforts across the Federal Government to enable Science and Technology, with participation from academia, industry, and international partners [72].

CHAPTER 3

TECHNOLOGY BACKGROUND

This chapter provides an in-depth background of the technologies used for this research project.

Natural Language Processing

Overview

Critical health information that can be used for surveillance purposes can be found in free-form text; however, the analyses of these data have inherent problems and should utilize structured data to successfully identify target population. The goal of natural language processing (NLP) is to classify, extract, and code information from free text data. Natural language processing has long been used to process text such as patient records and to discharge summaries. Over the past few decades, many research groups have been developing natural language process systems to aid in clinical decision support and research, quality assurance, and the automation of encoding free text data [73, 74].

Natural Language Processing (NLP) in the Biomedical Field

One important application area that has been ongoing for the past two decades as it relates to NLP is its application to the biomedical medical field [73, 75]. The Linguistic

String Project-Medical Language Processor (LSP-MLP) [76], which was developed out of the Linguistic String Project [77-79] at New York University, is one of the first large-scale medical NLP systems [75] and has been adopted for applications in French and German [80, 81]. LSP-MLP is aimed at enabling physicians to extract and summarize data, such as drug dosage and symptoms, to identify possible side effects of medications [82] .

Another system that is popular in the biomedical domain is the Medical Language Extraction and Encoding (MedLEE) system. This NLP system was developed, evaluated, and deployed by the Columbia Presbyterian Medical Center and is aimed at extracting textual information from clinical documents, such as cardiology reports, discharge summaries, pathology, and visit notes and more, then transforms the information into a structured and conceptual representation [83]. MedLEE allows clinical reports to be stored in a database that can be queried by physicians using controlled vocabularies. Although a large number of the reports processed using MedLEE is chest X-ray reports, it has been evaluated for a variety of tasks, including detecting patients with suspected tuberculosis [84]; identifying breast cancer [85], stroke [67], community-acquired Pneumonia [86], and healthcare-associated pneumonia in neonates [87]; assessing quality of care for cardiovascular [88]; automating coding of ICD-9 CM [89] and SNOMED [90]; and deriving comorbidities from text [91].

Since MedLEE, many institutions have developed numerous NLP systems for different biomedical tasks. Researchers from The University of Utah developed SymTextm [92, 93] which has been used for applications such as coding chief complaints into ICD-9 code [93], identification of pneumonia from chest X-ray reports

[94], and interpreting of chest radiograph reports for identifying the mention of a central venous catheter (CVC) [95]. KnowledgeMap [96] is another NLP system that was developed at Vanderbilt University. KnowledgeMap has been used for the extractions of medical concept from both clinical and educational documents [97]. Other research groups have created and evaluated NLP systems for processing biomedical information; for an in-depth review of these systems, please refer to the literature review conducted by Doan et al. [98].

Another popular NLP system in the biomedical domain is MetaMap [99], which was developed by researchers at the National Library of Medicine (NLM). MetaMap identifies biomedical concepts from free-form textual input and maps them to concepts from the Unified Medical Language System (UMLS) Metathesaurus [99-101]. For each inputted text, MetaMap breaks the words into phrases, classifies it into a semantic type, and then returns the concept unique identifier (CUI). Each matched phrase is ranked according to their calculated mapping strength [99]. MetaMap is widely by researchers in the biomedical community and has become the de facto standard.

Despite the wide use of natural language process tools in the clinical domain, there are few applications in the public health realm [102, 103]. However, natural language processing methods and techniques have been used in modern, hospital-based surveillance systems.

The Role of NLP in Public Health Surveillance Systems

Early detection of disease outbreaks has been a public health function and goal for many centuries [1] because it can lead to implementation of control measures and dramatically reduce morbidity and mortality rates for a population [2]. Early detection requires aberration detection algorithms to detect potential outbreaks and to facilitate rapid response. Many aberration detection algorithms require structured data to identify anomalous patterns for a particular time or location. The Handbook of Biosurveillance [104] defines structured data as “data in a format (e.g., relational database tables) that can be interpreted by a computer” [104]. Public health surveillance systems commonly use structure data, such as ICD-9 codes, to evaluate the performance of new surveillance systems; these are best known as gold standard data [104] . Such data include ICD-9 coded hospital discharge diagnoses and outpatient billing diagnoses. Another example of structured data used in today’s surveillance system is over the counter pharmacy sales (OTC) data, which counts the Global Trade Item Number or GTIN for each store [105]. These are just some of the structured data used for public health surveillance.

The US public health system is a large, distributed entity that captures a wealth of information. However, much of the data useful for public health surveillance are usually available in unstructured, free text format that can easily be read and understood by humans, but is difficult for computers to decipher [12]. Moreover, the analyses of these unstructured data have inherent problems and should use structured data to successfully identify target population. In addition, it is postulated that the ability to transform these data into a more structured format is ideal for surveillance systems because it provides greater sensitivity over specificity, allowing greater detection of events of interest [12].

Natural Language Processing and Death Certificates

In relation to this dissertation, little is known about using natural language processing methods and tools to automatically structure and code death certificates for disease surveillance purposes. Recently, Medical Match Master (MMM) [106] was developed to match unstructured cause of death literals to concepts (CUIS) and semantic types within the Unified Medical Language System (UMLS). The system was able to identify an exact concept identifier (CUI) from the UMLS for over 50% of cause of death literals, thus showing that the UMLS Meathesaurus contains adequate biomedical and public health-related concepts for processing death certificates. Therefore, for this research study, the natural language processing tool, MetaMap, was used to standardize and code the death certificates in the proposed death certificates pipeline (DCP). The Methods chapter (Chapter 4) provides a detailed description of how MetaMap can be used to process public health data.

Grid Technology

As we know it today, grid technology is touted to be the evolution and amalgamation of many development efforts in the computer science field that have been going on for many years. This section provides a brief historical background of grid computing and some existing tools and technologies for it. This is followed by a detailed discussion on grid technology in the biomedical domain.

Formal Definition of "Grid"

Over the years, there have been many definitions for "grids." Foster et al. first defined the concept of a grid as "controlled and coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations, VOs" [107]. In a grid environment, federated data and applications are pooled into virtual system allowing data owners to share data and applications while maintaining control [108]. This feature, which is commonly referred to as the "grid vision," is important for systems in the biomedical informatics domain, particularly public health informatics, because it provides officials/users the ability to have access to external distributed data.

Grid Computing in Biomedical Informatics

The notion of creating virtual organizations is appealing to many scientists, particularly academic and research organizations who are interested in taking advantage of unused resources. As a result, the field of biomedical informatics has widely utilized modern distributed systems using grid technologies [109] for research purposes. Grid-based computer technology is a distributed form of computing that provides clients with secure mechanisms to integrate disparate data and applications in a dependable, consistent, and inexpensive manner [109].   In a grid environment, federated data and applications are pooled into a virtual system allowing data owners to share data and applications while maintaining control [108]. This feature is important for systems in public health because it provides officials/users the ability to have access to external distributed data. Grid computing is widely applied in a field of biomedical informatics, particularly in translational clinical research. For example, the cancer Biomedical Informatics Grid (caBIG), which was a nationwide initiative led by the U.S. National Cancer Institute (NCI), is one of the first publicly available infrastructure to provide a federated approach to increase interoperability among research information systems [108]. The caBIG uses a grid infrastructure to connect disparate data and tools institutions, such as cancer centers, in the United States and worldwide [110].

Grid technologies applied to biomedical informatics can be classified into three categories from the biomedical informatics domain: clinical research (translational)/ bioinformatics, healthcare informatics, and public health informatics. Although the grid is applied to other biomedical informatics subdomains, the above classification highlights the subdomains in which grid computing is common and can provide insights into the advantages and disadvantages of using grid technologies for real-life applications.

Healthcare Informatics

As previously stated, grid technology allows the sharing of resources such as computer power, access to data, and/or analytic tools over the Internet. The groundbreaking white paper from the HealthGrid [111] initiative, which highlighted the requirements using grid technologies in biomedical healthcare, resulted in many clinicians and researchers promoting the advantages of deploying grid data and applications. Here we will provide a few descriptions of grid projects in the healthcare field with a focus on three common categories: healthcare imaging, computation analysis, and screening and detection of patients.

The GLOBUS-MEDICUS (Medical Imaging and Computing for Unified Information Sharing) project is an initiative created by a collaborative effort between University of South California and the Information Sciences Institute. The goal of this initiative is to share global clinical images through a seamless integration of Digital Imaging and Communication in Medicine (DICOM) standard protocol and devices, which are used widely in healthcare and medical research enterprises, into grid services [112]. In the grid, medical images became transparently available anywhere within a VO, comparable to a Regional Health Information Organization (RHIO) of hospitals or practices, and between VOs. The Globus MEDICUS grid was successfully deployed across 40 medical centers and it demonstrated the value of the grid model when applied to overcoming barriers to sharing clinical information [112]. VirtualPACS is another project aimed at providing clinicians access to distributed image databases; however, it goes a step further by providing users access to PACS and DICOM functionalities (such as query, retrieve, and submit) as if the images were stored in a centralized PACS system [113].

In Europe, researchers were developing the MediGrid, which was the development of a “grid middleware platform” with Globus Toolkit 4 as a basis to provide an easily accessible clinical environment for researchers to easily test and deploy new methods for image reconstruction [114]. The work resulted in a set of platforms of independent software tools which are able to read medical images, control the execution of computing intensive tomographic algorithms, and explore the reconstructed tomographic volumes. As a result, clinicians are able to manage, process, and visualize tomographic and ultrasound data from any geographic location.

Besides sharing and visualizing image data, researchers have also used grid technology for on-demand image analysis. Bagarinao et al. proposed the use of grid technology for the analysis of brain imaging data, such as functional MRI (fMRI) data that are computer resource intensive. To facilitate the analysis of fMRI datasets in a grid environment, a software package called BAXGrid was developed. This software provided researchers and medical practitioners the ability to perform onsite analysis of the generated data within seconds after data acquisition [115]. Another project, CardioGRID, is a system proposed by Khalil et al. [116] to provide cardiologists access to an on-demand detection algorithm for the identification of cardiovascular abnormalities. In addition, this system had the functionality to calculate heart rate, classify heart beats, and use both time and frequency domain of the heart rate variability to diagnose many cardiovascular diseases. The authors believed that the system would be useful in a wide range of patients from serious cardiac patients and aging populations to persons who would like periodic wellness monitoring.

Grid technology in healthcare informatics has been increasing in decision support and for cancer therapy and monitoring. For example, GridCAD was developed using the NCI Cancer Biomedical Informatics Grid architecture that allows the querying of both central and federated images for two purposes: 1) visualization and 2) use of grid enabled computer-assisted detection (CAD) algorithms for the detection of lung cancer [117]. Another example of using CAD algorithms in a grid environment is seen in a similar project, MammoGrid, which ran from 2002 to 2005, and was an initiative aimed at creating a European-wide database of mammograms in an effort to promote collaboration between European Union healthcare professionals. With MammoGrid, physicians not only have access to European wide mammograms but are able to run advanced algorithms on these images, including computer-aided detection analyses [118]. Another project whose aim was to use grid technology to help cancer patients is RadiotherpayGrid (RT-Grid). RT-Grid uses grid technology to reduce computation time of Monte Carlo simulations when used to calculate radiotherapy dosage. The reduction in computation time allows physicians to more quickly provide treatment to cancer patients and can help in complex cases where treatment plans are not as forthcoming [119].

Translational/Bioinformatics

Bioinformaticians use many relatively independent small computation tasks with thousands of relatively small independent tasks (high throughput computing) or use a suite of resources (from data to applications) to complete a specific tasks. For these reasons many grid applications in the bioinformatics domain aim at one of the following —grid enabling high-throughput applications or integrating biomedical applications, data, or services; the latter is highly relevant to this dissertation.

It is said that one of the best examples of using grid technology for high-throughput computing in bioinformatics is the Worldwide In Silico Docking On Malaria (WISDOM-I) [120] project. This was the first large-scale deployment of molecular docking application where it allowed scientists, who were in search for a drug that will combat malaria, to drastically increase the average number of drug compounds analyzed per hour. The project achieved 42 million dockings using 1700 computers distributed in 15 countries over a 2-month period; such computation would take approximately 80 years on one CPU. The authors later developed DIANE (Distributed Analysis Environment), an enhanced, lightweight version of WISDOM [121]. WISDOM DIANE was used to search potential drugs for possible variants of the avian flu virus (H5N1).

Sequence analyses are very typical processing applications in bioinformatics. One tool that is widely used by biologists for such analyses, Basic Local Alignment Search Tool (BLAST), is used for homology searches. Due to its popularity, there have been many implementations of a grid-based version of BLAST [122-128]. The grid-based version of BLAST focuses on 1) processing BLAST queries almost N times faster than if working with only one computer; 2) providing faster response time than the BLAST publicly available service; and 3) assembling of the results from distributed jobs.

The increased use of bioinformatics tools and data by biologists on a daily basis brought about the need to integrate the growing number of biological data and analytical services that are widely available on the web; bioinformatics workflow tools make such a task possible. The popularity of scientific workflow systems in the bioinformatics domain has resulted in many being developed within a short period of time. Examples include Taverna [129], Triana [130, 131], Tavaxy [132], Galaxy[133], DiscoveryNet [134], Pegasus [135], OMII-BPEL[136], YAWL; yet another workflow language [137], Kepler [138], Conveyor [139], Pegasys [140], Discovery Net [19,20], and OMII-BPEL [21]. An in-depth review of these tools can be found in [141] and [142] but perhaps one of the best known workflow applications in bioinformatics and highly recommended by the caGrid development team is Taverna [129]. Taverna facilitates quick building, running, and editing of workflows in a user-friendly interface, and has the functionality for users to integrate published web services [129]. Taverna was an integral part of the myGrid e-Science initiative because it provided developers informatics-related functionalities that are lacking in its competitors such as metadata repositories and ontology–driven search tools.

Public Health Informatics

Grid computing applied to the public health domain has very few applications; however, in the past decade, grid for public health purposes has been on the rise. Several prototypes for grid-based surveillance systems have been developed in Europe (e.g., grid technology for avian flu and cancer surveillance [143, 144]), Asia (access grids to aid SARS patients) [145], and South America (IntegraEPI) [146]. In the US, informaticist at the National Center for Public Health Informatics (NCPHI) at the CDC has promoted the idea of a public health grid that can be used to create a public health virtual organization to connect many public health entities such local and state public health departments, federal agencies, hospital and commercial lab systems, consumers, providers, and more.

NCPHI efforts led to publications demonstrating the potential value of grid technology to exchange information during the 2009 H1N1 pandemic [147].

The problems addressed using grid technology in the past are similar to those found in public health. Grid technology would allow a robust technology infrastructure for the exchange of data and resources across public health departments [148].

Overview of Proposed Solution

Using the technology described above, this research project aims at restructuring the current Mortality surveillance workflow. The three key stages in this workflow correlates to three key challenges faced by public health surveillance systems; for this reason, mortality surveillance workflow is used to demonstrate the feasibility of leveraging informatics techniques to advance and enhance surveillance processes; Figure 3.1 depicts the mortality surveillance workflow and identifies the proposed process.

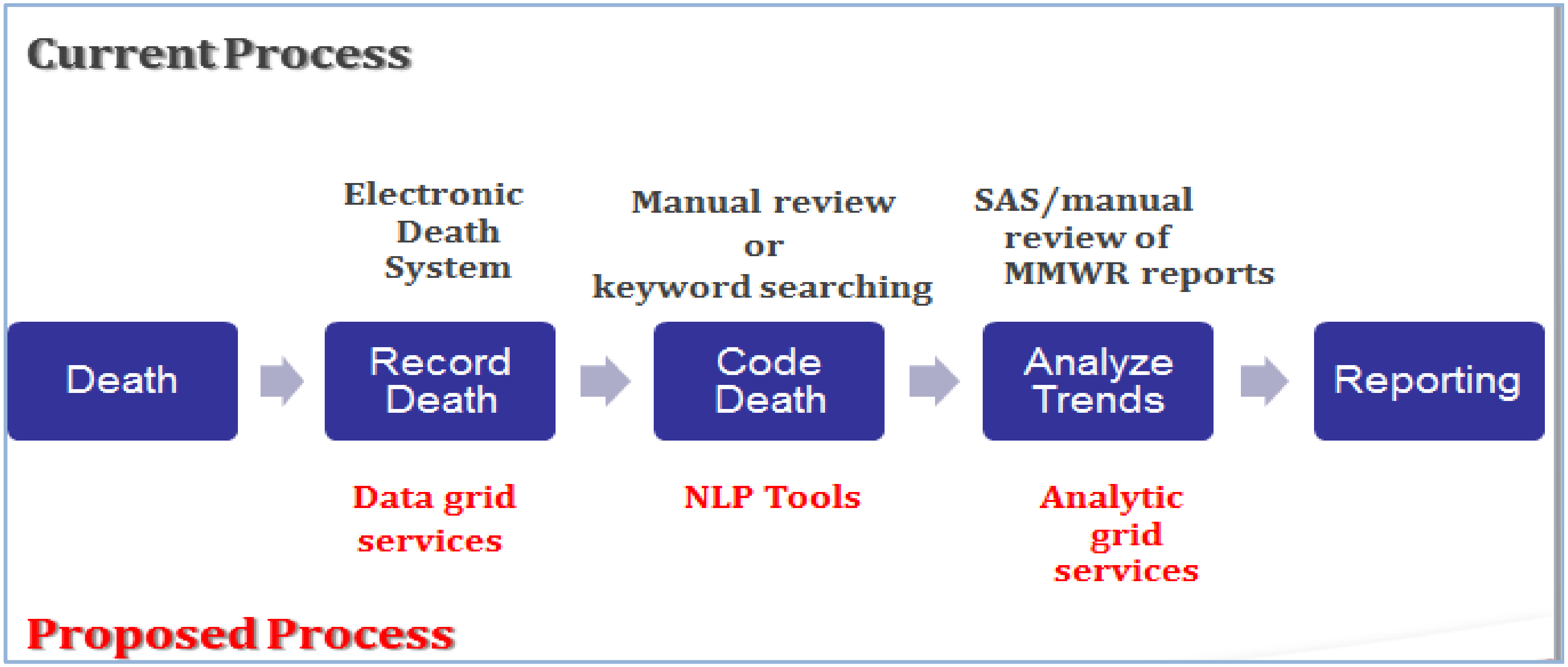


Figure 3.1: Mortality surveillance workflow current vs proposed process

CHAPTER 4

METHODS

Introduction

This chapter focuses on the experimental portion of the research where the core challenges identified in this study are investigated. The research was conducted in two stages. Phase one of the study involved the use of natural language processing methods to determine if this technique can be used to standardize and code the free text found on death certificates and used the standardized output for the automatic identification of pneumonia and influenza (P-I) deaths. Although this phase of the study is a small segment of the overall research project, it has the potential to demonstrate how unstructured public health data can be effectively used for surveillance purposes. Phase two of the study consisted of a series of feasibility studies to demonstrate the ability to share, process, and analyze public health data in a grid environment.

In this chapter, the background information about research approaches and methods used throughout the research is presented. Following this, a description of the instruments, sampling procedures, data collection processes and statistical analyses used to implement the different phases of the study is discussed. Lastly, an evaluation/assessment of the methods used for the different phases of the research project is also provided.

Research Philosophy and Approaches

Different research paradigms and methodologies are based on varying worldview and/or philosophical foundations [149]. Therefore, it is imperative that a researcher have a clear understanding of the research paradigm issues that guide and inform his/her research approach because each research paradigm, in turn, is implemented by associated methodological approaches and strategies [149]. In biomedical informatics research, Friedman et al. [150] noted that the most common philosophical approaches used are objectivist (positivism) and the subjectivist (interpretive) approach.

The objectivist philosophical position is the oldest and the most widely used of the two research approaches [150, 151]. Although there are many versions of objectivism, the overarching idea of this philosophical notion is that an objective reality exists external to human beings. Specifically to biomedical informatics, a key perspective of the objectivist paradigm is the objective assessment of clearly defined variables, usually measured quantitatively. The objectivist approach provides the theoretical basis for quantitative research. The first phase of this research project focuses on statistical measurements, thus we will follow the objectivist approach [150]. A quantitative approach employing many elements of measurement studies provided analysis and verification of the death certificate pipeline and rendered accurate measurement of the instruments under investigation.

The second phase of this project, which contributes to the majority of the project, however, is with the subjectivist approach. Subjectivism, by contrast, is based on the logic of interpretation in hopes of finding new interpretations and/or underlying meanings [152]. Although this paradigm of research emanated from social sciences, recently, the subjectivist viewpoint has received attention in the field of biomedical informatics because developers and implementers realize the impact of social, people, and organizational issues of creating or implementing health-related systems [153]. There are several models of the subjectivist paradigm, one of which is grounded theory [154]. Phase two, the primary component of this project, follows a grounded theory approach. In this methodological approach, the researcher concentrates on gaining ideas or insights through secondary or exploratory research, such as pilot/feasibility studies [155].

Feasibility Studies

Feasibility studies aim to investigate and evaluate the use of the proposed methods and techniques to determine how well they perform and if they present a viable solution for the problem being investigated [156]. The decision to use this study method approach for the second phase of the research project was driven by several factors, which are listed below:

- Feasibility studies allow researchers an opportunity to test a technique to a problem that has limited or no information, and to determine if the method is an applicable solution to the problem [156].
- The results from a feasibility study provide insight of the technique to those who are considering the implementation of similar projects. Such insight is invaluable for the researcher who may have identified and/or addressed any possible problems or difficulties that may arise [156].

Feasibility studies are exploratory by nature, which can be a major disadvantage. The results from this type of study only suggest options to consider, and cannot be used as conclusive evidence that the implementation of similar projects would be easily reproducible or advisable [157]. However, this limitation was not a critical factor for this research project because the purpose of the feasibility studies undertaken were aimed at providing guidance and/or suggestions for similar future projects.

Research Methods

We used a mix methods approach to address the research questions for this dissertation. This approach is supported by Brewer et al. [158] who state that it may be advantageous to use multiple methodological approaches to address the research problem. The following sections expand upon the implementation of each phase.

Methods for Phase One

Introduction

The aim of the first part of the project was to demonstrate the feasibility of creating a death certificates pipeline, which includes a natural language processor and detection rules, for the coding of death certificates and identification of potential diseases; for the purpose of this study, the identification of pneumonia and influenza cases was selected. Also, we aimed to demonstrate that the pipeline's accuracy is comparable to existing methods. For this phase of the study, a measurement study approach was identified as the most fitting way in which to accomplish this goal.

Measurement Study

Measurement studies are used to help researchers determine the accuracy an attribute of interest can be measured "in a population of objects." It is acknowledged in the biomedical informatics domain that such studies are outstanding for comparing a new instrument developed to a “gold/reference standard” [150]. A gold/reference standard is considered the “truth” about the condition of a task. All domain might not have an absolute “truth”; therefore, some studies use the best approximation of the “truth” that is available to the investigator [150].

Data Collection

We obtained over fourteen thousand electronic death records from the Utah Department of Health (UDOH) for the period January 1, 2008 to December 31, 2008. The records included a field describing the disease or condition directly leading to death and any antecedent causes, comorbid conditions, and other significant contributing conditions.

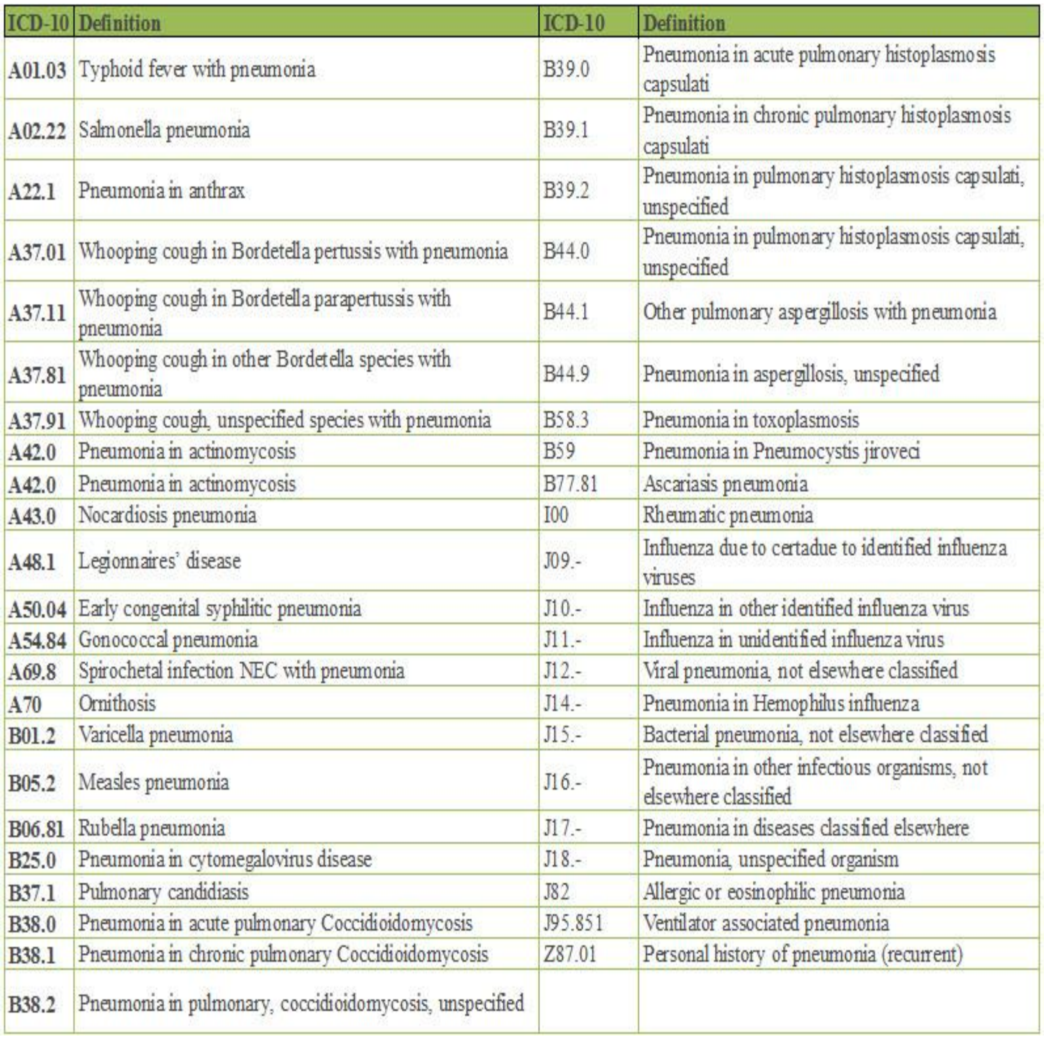
All death records used in this study have been processed by the National Center for Health Statistics (NCHS) using the Mortality Medical Data System (MMDS). Through a suite of programs, the MMDS software automates the coding of cause of death literals into International Classification of Diseases Tenth Revision (ICD-10) and selects the underlying and up to 20 multiple cause of death codes based on the World Health Organization coding rules [159]. The software generates two different sets multiple cause of death codes for tabulation: entity-axis and record-axis codes. The entity axis codes contain ICD-10 codes representing all conditions listed on the death certificate and preserves the order in it appeared. The other, record axis codes, contains a set of ICD-10 codes that best describe the overall medical certification portion of the death certificate. NCHS only provides the record axis codes to the health departments; therefore, this type of code was used for this study. It is important to note that of the 2.3 million yearly deaths, 80–85 percent are automatically coded through MMDS [160]. As a result, some of the codes used in this study were also manually coded into ICD-10 by a nosologist, a medical classification specialist. For more information on the coding process, refer to Appendix A.

For our study, we randomly selected 45% of the 14,440 records (6,450). The death records in this study were previously coded by the Mortality Medical Data System (MMDS) software or manually coded by a nosologist at the National Center for Health Statistics into ICD-10 codes. However, this information was only used as a posteriori to measure the quality of the automatic coding.

Pneumonia and Influenza (P-I) Deaths Case Definition

We decided to apply the CDC's operational case definition for pneumonia and influenza deaths; this definition was defined by a CDC epidemiologist staff through personal communication. As a result, for our study, the operational definition for a pneumonia or influenza death is: an influenza death is defined as all types of influenza with the exception of deaths from PARAINFLUENZAE VIRUS infection and HAEMOPHILUS INFLUENZAE infection. A pneumonia death is defined as deaths from all types of pneumonia with the inclusion of pneumonia due to parainfluenzae virus and pneumonia due to H. influenza. Pneumonia exclusion includes pneumonitis (related ICD-10 codes: J67-J70 and J84.1), aspiration pneumonia (related ICD-10 codes: J69.-, O29, O74.0, O89.0, and P24.-), and pneumonia due to pneumococcal meningitis (related ICD-10 codes: G00.1 and J13).

Using the operationalized definition from the CDC, pneumonia- and influenza-related deaths were defined as one the codes listed in Figure 4.1. The related codes were selected by an extensive, manual review of the ICD-10 2007 instruction manual [161].



ICD-10	Definition	ICD-10	Definition
A01.03	Typhoid fever with pneumonia	B39.0	Pneumonia in acute pulmonary histoplasmosis capsulati
A02.22	Salmonella pneumonia	B39.1	Pneumonia in chronic pulmonary histoplasmosis capsulati
A22.1	Pneumonia in anthrax	B39.2	Pneumonia in pulmonary histoplasmosis capsulati, unspecified
A37.01	Whooping cough in Bordetella pertussis with pneumonia	B44.0	Pneumonia in pulmonary histoplasmosis capsulati, unspecified
A37.11	Whooping cough in Bordetella parapertussis with pneumonia	B44.1	Other pulmonary aspergillosis with pneumonia
A37.81	Whooping cough in other Bordetella species with pneumonia	B44.9	Pneumonia in aspergillosis, unspecified
A37.91	Whooping cough, unspecified species with pneumonia	B58.3	Pneumonia in toxoplasmosis
A42.0	Pneumonia in actinomycosis	B59	Pneumonia in Pneumocystis jiroveci
A42.0	Pneumonia in actinomycosis	B77.81	Ascariasis pneumonia
A43.0	Nocardiosis pneumonia	I00	Rheumatic pneumonia
A48.1	Legionnaires' disease	J09.-	Influenza due to certadue to identified influenza viruses
A50.04	Early congenital syphilitic pneumonia	J10.-	Influenza in other identified influenza virus
A54.84	Gonococcal pneumonia	J11.-	Influenza in unidentified influenza virus
A69.8	Spirochetal infection NEC with pneumonia	J12.-	Viral pneumonia, not elsewhere classified
A70	Ornithosis	J14.-	Pneumonia in Hemophilus influenza
B01.2	Varicella pneumonia	J15.-	Bacterial pneumonia, not elsewhere classified
B05.2	Measles pneumonia	J16.-	Pneumonia in other infectious organisms, not elsewhere classified
B06.81	Rubella pneumonia	J17.-	Pneumonia in diseases classified elsewhere
B25.0	Pneumonia in cytomegalovirus disease	J18.-	Pneumonia, unspecified organism
B37.1	Pulmonary candidiasis	J82	Allergic or eosinophilic pneumonia
B38.0	Pneumonia in acute pulmonary Coccidioidomycosis	J95.851	Ventilator associated pneumonia
B38.1	Pneumonia in chronic pulmonary Coccidioidomycosis	Z87.01	Personal history of pneumonia (recurrent)
B38.2	Pneumonia in pulmonary, coccidioidomycosis, unspecified		

Figure 4.1: Pneumonia and influenza ICD-10 codes used in this study

Study Procedures

The Death Certificates Pipeline, DCP, was created to code death certificates and to identify pneumonia- and influenza-related cases. The pipeline consisted of two major components. The first component was a natural language processor tool to code the death certificates; for this, we used the National Library of Medicine MetaMap [99] tool. The second component was the definitional rules engine; the rules were applied to MetaMap's output to identify our cases of interest. The DCP steps, which are depicted in Figure 4.2, includes: preprocessing of the data, a natural language process tool to code and standardize the data, extraction of coded data, and the detection of pneumonia and influenza cases. A detailed explanation of each step is provided in the next section.

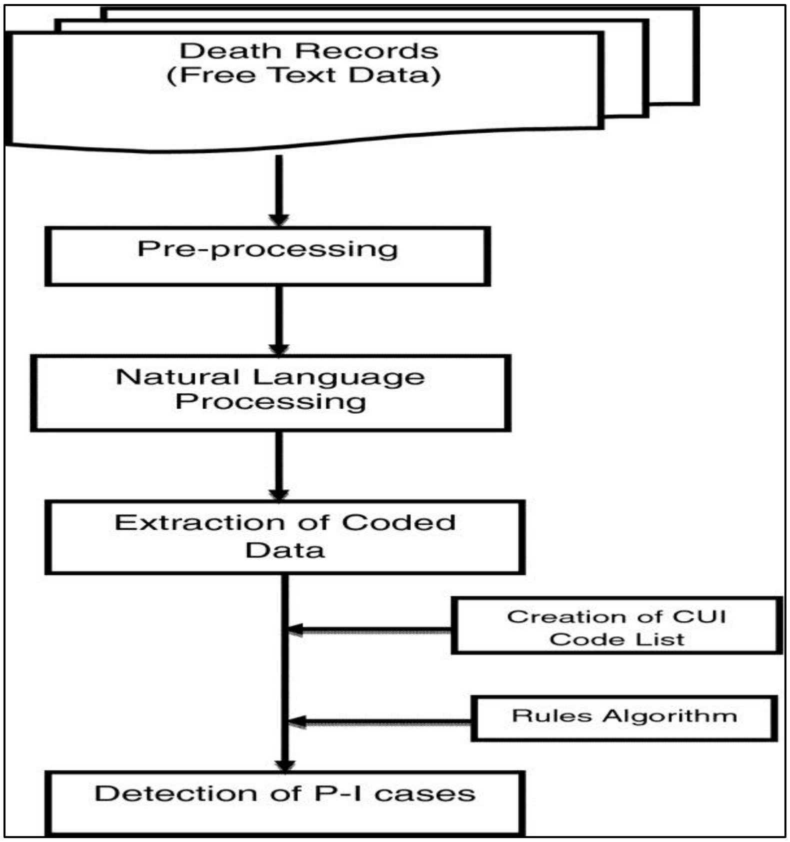


Figure 4.2: Death Certificates Pipeline (DCP) workflow diagram

Step 1: Preprocessing

A common error found on death certificates is spelling errors; therefore, all death certificates used in this phase of the study were processed through a spell checker to identify misspellings. For MetaMap, which uses the Unified Medical Language System (UMLS) [162], the National Library of Medicine recommends the use of the spell checker tool called GSPELL [163, 164]. However, we chose to use ASPELL [165] to satisfy our spell checking needs. This decision was primarily based on the findings of an evaluation by Cromwell et al. [164], which showed ASPELL performing better than GSPELL in three areas: (1) correct word was ranked as number one; (2) correct word was ranked in the top ten; and (3) correct word was found by the spell checker [164].

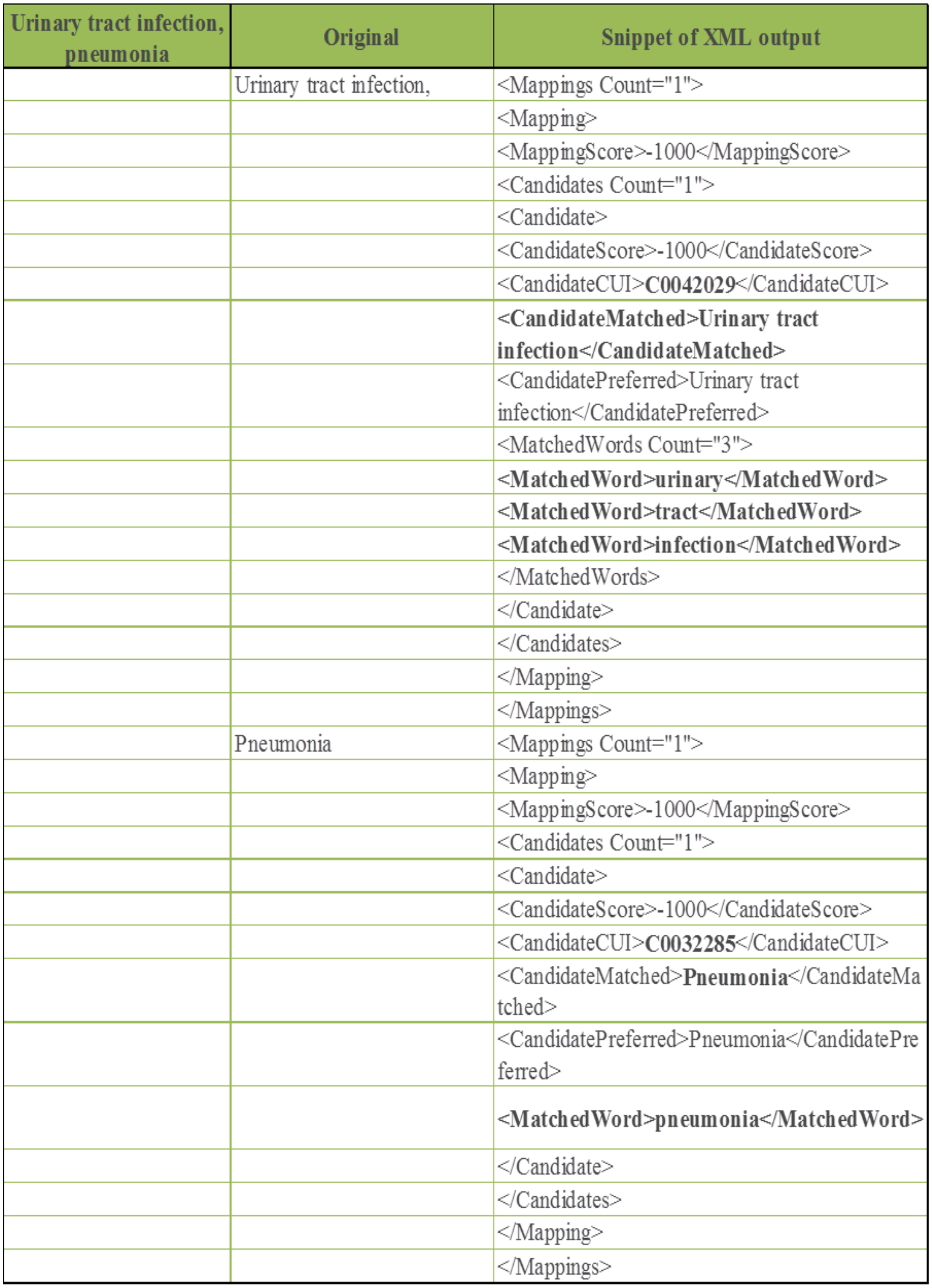
The next preprocess step used the scripting language PERL (http://www.perl.org) to prepare the death data for MetaMap. During this step, non-ASCII characters were removed from the death literal; this was a required for MetaMap to process the data.

Step 2: Natural Language Processing

As noted earlier, MetaMap was the natural language processing tool selected to code and standardize the 6450 death certificates. Figure 4.3 shows an example of sample death literals and its associated XML output from MetaMap. Text bolded in the output represents MetaMap code, known as CUI, and its corresponding death literal.

Step 3: Extraction of Coded Data

MetaMap's XML output of the death literals was processed through a PERL script. The script extracted the processed death literal text and its corresponding CUI(s). The information was then outputted to a comma-separated values (CSV) document.



Urinary tract infection, pneumonia	Original	Snippet of XML output
	Urinary tract infection,	<Mappings Count="1">
		<Mapping>
		<MappingScore>-1000</MappingScore>
		<Candidates Count="1">
		<Candidate>
		<CandidateScore>-1000</CandidateScore>
		<CandidateCUI>**C0042029**</CandidateCUI>
		<CandidateMatched>Urinary tract infection</CandidateMatched>
		<CandidatePreferred>Urinary tract infection</CandidatePreferred>
		<MatchedWords Count="3">
		<MatchedWord>urinary</MatchedWord>
		<MatchedWord>tract</MatchedWord>
		<MatchedWord>infection</MatchedWord>
		</MatchedWords>
		</Candidate>
		</Candidates>
		</Mapping>
		</Mappings>
	Pneumonia	<Mappings Count="1">
		<Mapping>
		<MappingScore>-1000</MappingScore>
		<Candidates Count="1">
		<Candidate>
		<CandidateScore>-1000</CandidateScore>
		<CandidateCUI>**C0032285**</CandidateCUI>
		<CandidateMatched>**Pneumonia**</CandidateMatched>
		<CandidatePreferred>Pneumonia</CandidatePreferred>
		<MatchedWord>pneumonia</MatchedWord>
		</Candidate>
		</Candidates>
		</Mapping>
		</Mappings>

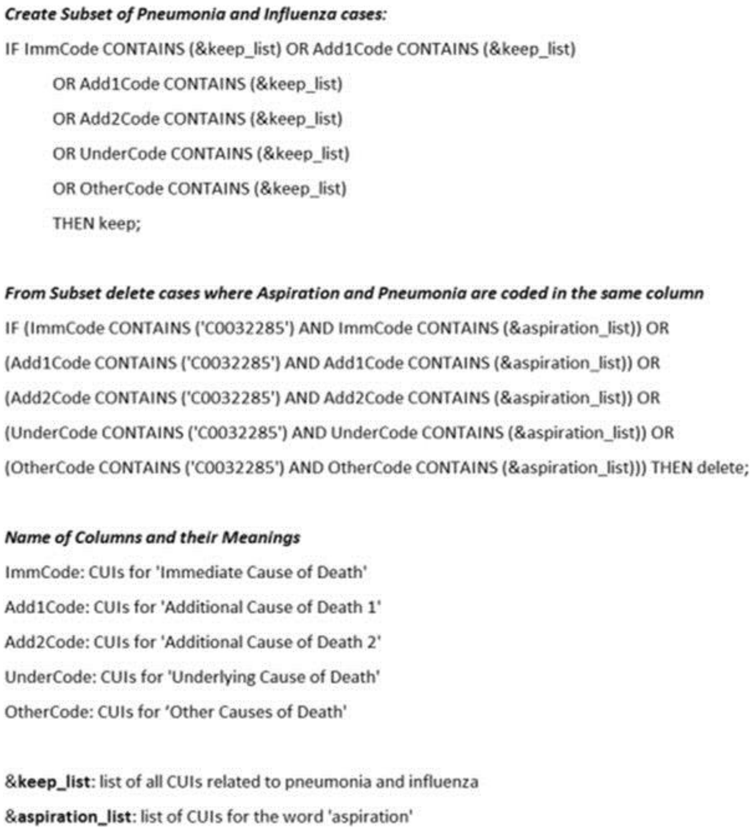
Figure 4.3: Original text and its corresponding MetaMap output

Step 4: Identification of P-I Deaths

The identification of pneumonia- and influenza-related cases from the standardized death data involved two steps: 1) mapping selected pneumonia- and influenza-related ICD-10 codes to the appropriate MetaMap CUI codes and 2) using the MetaMap CUIs to create a rules-based algorithm to identify cases of interest. To select the appropriate CUIs, we created a "CUI code list" that represents all the CUI codes of interest. Several steps were taken to create this list. First, a subset of the UMLS 2010 AB database was generated using the Metamorphosys [101] tool provided by the National Library of Medicine, NLM. The UMLS Metathesaurus database includes many terminology/vocabulary sources, as a result, we adopted the procedure used by Riedl et al. [106] to select the most appropriate terminology/vocabulary sources relevant to our aims. This method selected every level 0 sources plus SNOMED-CT, a level 9 source. Death certificates are coded into ICD-10 codes; therefore, for our configuration, we decided to also include sources such as the National Center for Biotechnology Information (NCBI) taxonomy, which included ICD-10 codes. Two tables within the UMLS schema were primarily used to create the "CUI code list." The first table, MRCONSO, contains a unique row for each lexical variant of a given concept. The second table, MRREL, contains information about the relationship among concepts. The configuration produced 6,862,110 rows in MCRNOSO and 23, 467,822 rows in MRREL.

To map the selected pneumonia and influenza ICD-10 codes (Figure 4.1) to MetaMap CUIs, three queries were performed on the subset produced above. The results from the queries were combined and all duplicates were removed. A final query was then run to ensure that only the ICD-10 codes in Figure 4.1 were included in the final "CUI code list." This produced 241 distinct concept identifiers (CUIs) related to pneumonia or influenza. These codes were used to develop the rules to identify pneumonia and influenza cases (Figure 4.4).

The coded and standardized data produced by MetaMap were accessed by rules created in SAS; these rules were used to identify the pneumonia and influenza cases in the 6450 death records. The rules algorithm uses If-Then and Boolean operators (And, Or, Not) to create a rules chain (Figure 4.4). To identify pneumonia and influenza cases, the algorithm looks at each cause of death field within a death record (e.g., Immediate Cause of Death, Underlying Cause of Death, Additional Cause of Death, etc.) to flag relevant records.



Create Subset of Pneumonia and Influenza cases:

```
IF ImmCode CONTAINS (&keep_list) OR Add1Code CONTAINS (&keep_list)
        OR Add1Code CONTAINS (&keep_list)
        OR Add2Code CONTAINS (&keep_list)
        OR UnderCode CONTAINS (&keep_list)
        OR OtherCode CONTAINS (&keep_list)
        THEN keep;
```

From Subset delete cases where Aspiration and Pneumonia are coded in the same column

```
IF (ImmCode CONTAINS ('C0032285') AND ImmCode CONTAINS (&aspiration_list)) OR
(Add1Code CONTAINS ('C0032285') AND Add1Code CONTAINS (&aspiration_list)) OR
(Add2Code CONTAINS ('C0032285') AND Add2Code CONTAINS (&aspiration_list)) OR
(UnderCode CONTAINS ('C0032285') AND UnderCode CONTAINS (&aspiration_list)) OR
(OtherCode CONTAINS ('C0032285') AND OtherCode CONTAINS (&aspiration_list))) THEN delete;
```

Name of Columns and their Meanings

ImmCode: CUIs for 'Immediate Cause of Death'
Add1Code: CUIs for 'Additional Cause of Death 1'
Add2Code: CUIs for 'Additional Cause of Death 2'
UnderCode: CUIs for 'Underlying Cause of Death'
OtherCode: CUIs for 'Other Causes of Death'

&**keep_list**: list of all CUIs related to pneumonia and influenza
&**aspiration_list**: list of CUIs for the word 'aspiration'

Figure 4.4: Rules applied to MetaMap's output to extract pneumonia and influenza cases

Comparison of Methodologies

MMDS is frequently considered by practitioners as the "gold standard" for the processing and coding of death certificates in the US and many other countries. Therefore, the codes produced by this system were used as the "reference standard" to determine the performance of two methods: a) our automated detection system, DCP and b) the current technique used at UDOH, keyword searching.

To identify pneumonia and influenza cases using the keyword search method, we followed the current process used by the Utah Department of Health where all the cause of death fields were scanned for the text strings "PNEUMONIA" OR "INFLUENZA," while excluding records that contained the words "ASPIRATION PNEUMONIA," "PNEUMONITIS," "PNEUMOCOCCAL MENINGITIS," "HAEMOPHILUS INFLUENZAE" OR "PARAINFLUENZAE VIRUS."

To measure the performance of both methods against the reference standard, we needed to specify what constituted a match. Each death record is associated with a unique number; therefore, we considered a match if the unique identifier was identified by the comparator (the DCP and/or keyword searching) and also found by the reference standard.

Statistical Analysis

Three common standard measures were used to evaluate the performance of the two methods, DCP and keyword searching: recall, precision, and F-measure. The formula used to calculate these values are as follows:

- Precision = TP/(TP + FP) (1)
- Recall = TP/(True Positives + False Negatives) (2)

- F-measure = 2 *(P R/ P + R) (3)

For the purpose of this study, to calculate the values above, we had to classify flagged deaths as true positive, false positive, or false negatives. The definitions for each are:

- True Positive: a death that the comparator and the reference standard both classified as a P-I related death
- False Positive: a death that was classified as a P-I related death by the comparator, but not by the reference standard.
- False Negative: a death that was not classified as a P-I related death by the comparator, but was classified as a P-I death by the reference standard.

Cohen's Kappa was also used to assess agreement with the reference standard. Fisher's exact test was used to detect the significance of the differences between recall and precision between the two comparators. Finally, McNemar's test was used to determine if there was a statistical difference between the two methods. All calculations were performed in R [166].

To calculate the McNemar's test value, a confusion matrix was created where A is the number of times both methods have correct predictions; B is the number of times method 1 has a correct prediction and method 2 has a wrong prediction; C is the number of times method 2 has a correct prediction and method 1 has a wrong prediction; D is the number of times both methods have incorrect predictions.

For more information regarding the methods for this phase of the study, refer to the published article by Davis et al. [167] or Appendix A.

Methods for Phase Two

Introduction

The core aim of phase two was to use grid technology to explore and develop a decentralized architecture for sharing death certificate records, and to allow data analysis using an open source statistical software package in a grid environment. As such a study has not been done before, several methods were examined to allow us to choose the most appropriate method for this phase of the study. We decided to use the feasibility studies approach because they are appropriate when initial ideas need to be further developed [157].

Defining Elements for the Feasibility Study

Information gathered from the literature and from the researcher and mentor's background knowledge motivated the decision to include the element below for the development of the feasibility study:

- Data for Mortality Surveillance
- Applications for Mortality Surveillance Purposes
- Grid Environment
- Type of Grid Service

Step 1: Select Data for Mortality Surveillance

We used the 62,181 death records from the Utah Department of Health for this phase of the study.

Step 2: Select Applications for Mortality Surveillance

We considered the following criteria when selecting applications for this feasibility study:

- Level of acceptance and/or use by the research community
- Level of acceptance/ interest and/or use by the public health community
- Low cost for acquiring the application
- Level of maturity of the methods utilized in the analytical tool
- Ability to enhance current mortality surveillance methods
- Ability to bring novelty to a step or steps in the mortality surveillance workflow

After evaluating several commonly used tools in the biomedical informatics and public health domain, the following two applications were selected: MetaMap and R. MetaMap was selected because it met the selection criteria and has already been successfully used during phase one of the research.

R also met the selection criteria. It is an open source programming language for performing statistical analyses and creating basic to advanced visualizations [168]. It is widely used among statisticians and biologist for analyzing large datasets. The public health domain is faced with ongoing funding cuts and budgetary restrictions thus hindering their ability to purchase commercial products and applications. As a result, there is an increased interest in the public health domain for open source software for the collection, analysis, and visualization of public health-related data. As a result, the public health community has been pushing the use of R in state and local health departments and has been offering the public health workforce, particularly public health practitioners, hands-on training in R. Moreover, evaluation of the tool's offering to public health surveillance showed that in the Spring of 2013, there were approximately 20 packages that epidemiologists could use for various analyses of epidemiological data. For example, cluster detection tools are available through the use of the DCluster [169] or SpatialEpi [170] packages, and outbreak visualization can be accomplished using epitools [171]. The "surveillance" package, which implements a variety of commonly used algorithms for the detection of aberrations in routinely collected surveillance data, was most relevant for this research.

The R surveillance package is intended to provide users an open source software for outbreak detection and the visualization, modeling, and monitoring of routinely collected surveillance data [172]. The "surveillance" package functionality is divided into two categories: prospective change-point (aberration) detection algorithms and retrospective modeling algorithms [172]. Traditional public health aberration algorithms such as the function "cdc," which implements the approach described in Stroup et al. [173] and the function "farrington" implementing the statistical algorithm developed by Farrington et al. [174] can be found in the prospective algorithms functionality. Also, functions such as "cusum" [175], "rogerson" [176], and "glrnb" [177] provide a more statistical process control oriented approach.

The retrospective functionality allows the modeling of time series using either of the two functions: "algo.hhh," which implements the Held et al. [178] and Paul et al. [179] branching process approach, and "algo.hmm," which implements the hidden Markov model approach described in Le Strat and Carrat [180]. Also, "algo.twin"' uses the Held et al. [181] two-component endemic and epidemic approach. A list of the algorithms available in the surveillance" package can be found in Appendix B.

Step 3: Select Grid Environment

There are many grid environments available for both research and operational purposes; therefore, the grid environment selected has to be appropriate. For this particular study, caBIG [182] was chosen as the grid environment of choice for the following three reasons:

1. Its recognition for providing the appropriate environment for creating and deploying biomedical informatics and/or public health data and/or applications as grid services.
2. The availability of caBIG's extensive suite of tools to facilitate the creation and/or deployment of grid services.
3. The use of caBIG infrastructure to facilitate the querying of disparate data sources.

Step 4: Selecting Type of Grid Service

caBIG grid service are divided into two categories: data resources, which are used to expose data as data services; analytical resources, which are used to expose analysis applications as analytical services. As noted earlier, the purpose of phase two of the study is to demonstrate the feasibility of using grid technology to create a decentralized architecture for surveillance purposes, specifically mortality surveillance; therefore, for this project, both types of grid services were developed. The development of both services will allow the development of real-world scenarios for proof-of-concept validations.

Creating caGrid Data and Analytical Services

Creating caGrid Data Services

The caBIG team has developed a suite of tools for creating caGrid [183] data services based on the Unified Modeling Language (UML) methodology [184]. In this methodology, detailed description of data elements is used to describe abstract concepts through a series of steps. The Enterprise Architect [185] software was used to create the necessary UML models based on the recommendations in the caBIG developer documentation.

Implementation of caGrid data services requires the creations of two UML models: the Logical Model and the Data Model. The Logical model can be described as a representation of expert knowledge of the domain under study; for example, a logical model based on mortality surveillance can represent concepts such as “Name” and “Cause of Death” as classes that consist of attributes and relationships. On the other hand, the data model represents the database schema of the logical model where the logical model concepts/classes are represented by table and attributes.

After creating the data model, the elements were mapped to their corresponding elements found in the logical models; this annotation process allows developers to achieve caBIG silver-level compatibility. For this project, the annotation was done manually, which is the traditional approach. However, we do understand that this process can be error prone; therefore, the caAdapter tool [186] was used to visualize and validate the structure and mapping specifications, as well as to aid in finding any inconsistencies in the UML Model. This tool provides users the ability to automate and validate the mapping process, and also provides a component-based architecture to support message development and report using standard data formats. Once validated, the tool allows the annotations of the created UML model to be stored in the XML Metadata Interchange (XMI) representation.

Syntactic and semantic interoperability between resources is achieved by the caBIG team by using a model driven architecture approach. Therefore, after creating the UML models, we used the National Cancer Institute (NCI) caCore (Cancer Common Ontologic Representation Environment) SDK, which was built on the principles of Model Driven Architecture (MDA), n-tier architecture and common API for data access, to generate a "caCore like system" [187]. This system was used to create a caGrid data service that supports the use of existing caCore SDK-generated systems as a backend data source [187].

The next step for this caGrid data service creation involved the use of caGrid Introduce toolkit to generate and deploy the data services using the caCORE-generated system as the bases for the data service. The data services can be using the caGrid query languages: CQL is used to retrieve data from caGrid data services, and "DCQL," a distributed query language that is used for federated query processing.

Creating caGrid Analytical Services

To create the analytical services (R and MetaMap) for this study, we used an extension of the caGrid Introduce toolkit [188], grid Remote Application Virtualization Interface (gRAVI) [189]. The gRAVI toolkit allows developers to seamlessly wrap and deploy applications as a Globus compliant grid service. Before using gRAVI to create the analytical service, we first had to determine the parameter set that is required to execute both R and MetMap from the command line. Once this was established, gRAVI was used to “wrap” the R and MetaMap command line interface into a grid service. The service was then deployed and invoked using the respective grid service client for R and MetaMap, which was automatically created by Introduce and gRAVI.

For a detailed description of the methods used to create the analytical services for MetaMap and R, refer to the conference articles “Implementing public health analytical services: Grid enabling of MetaMap” [190] and “A Grid Based Approach to Share Public Health Surveillance Applications: The R Example” [191], or Appendices C and D, respectively.

Workflows

Like many scientific domains, public health analysis is increasingly data and process driven. It is noted that such analysis is well modeled by scientific workflows that orchestrate multistep procedures in a unified manner. Consider the example of the current steps to identify pneumonia and influenza deaths; a workflow can be created to automatically standardize, code, and identify these related deaths. Moreover, the workflow can be expanded to include operations to find different diseases or combinations of events and then be reported through notifications or published to other users via a website. Workflows can be used to integrate web or grid services. Despite a variety of workflow engines available, the caGrid team has proposed the use of Taverna [129], an open source workflow workbench that allows the creation, modification, and execution of workflows through a workbench graphical user interface (GUI), as its workflow engine for the following reasons:

1. Seamless integration with web service technology
2. Supports various web services

3. Includes text manipulation services
4. Plugin architecture that allows the integration of third-party extensions, such as the caGrid Workflow Toolkit that allows users to easily create workflows using caGrid services
5. Broad user acceptance within the biomedical informatics community
6. Relatively easy for developers to build and execute of scientific workflows
7. Extensibility

Mortality Surveillance Workflow

Traditionally, the mortality surveillance workflow includes manual processes, which may only be focused on a limited set of diseases and jurisdictions. For example, as noted in Chapter 1, the CDC's 122 City Pneumonia and Influenza mortality surveillance system publishes the number of pneumonia and influenza deaths for certain cities each week. The current process at the Utah Department of Health is shown in Figure 4.5.

While the general processes may be similar in different states, specific methods for each process may vary by state. For example, epidemiologists in each state use a variety of methods for case finding, and they may use many scripts run locally by an individual to analyze the findings. These practices make it difficult for standardizing the identification of pneumonia and influenza deaths across different states, or easily accessing counts from other states or over time.

For this study, the mortality surveillance workflow is constructed from the caGrid services described in the previous section. Each service is registered; as a result, they are discoverable from Taverna.

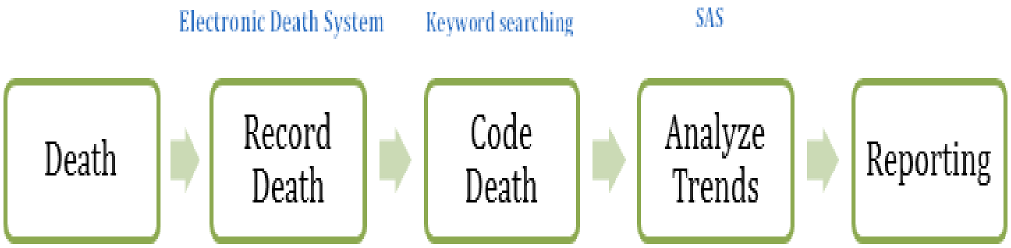


Figure 4.5: Current mortality surveillance workflow

Evaluation of Phase Two

When conducting a feasibility study, it is imperative to perform an evaluation. Therefore, for the evaluation of this phase of the study, we have identified “validation through proof-of-concepts” to guide our assessment. Through a series of use case scenarios, four proofs-of-concept studies were carried out. The methods for these uses are tied to the ‘Results’ of creating the grid services; therefore, a detailed discussion of the methods for these use cases is be provided in the ‘Results’ chapter (Chapter 5).

1. Use Case 1: Access and Integrating Public Health Data
2. Use Case 2: Access to Natural Language Processing Tools
3. Use Case 3: Deploy Commonly Used Public Health Surveillance Algorithms in a Grid Environment
4. Use Case 4: Access to Public Health Data and Informatics Services

Ethics approval was not required for this study; identifiable information was excluded from the study data.

Summary

This chapter described the research methodologies used to address the research questions identified for this dissertation. A measurement study was conducted to provide insight on how death certificates can be automatically standardized and coded. Building on that phase of the study, feasibility studies were conducted to demonstrate the transfer of death data and applicable applications to a grid environment.

CHAPTER 5

RESULTS

Introduction

The research methodologies described in Chapter 4 were categorized into two distinct phases, each with its own purpose and procedures. The results and finding obtained by the two research phases are presented in this chapter.

Results for Phase One

Processing Time of the Data

The 6450 death records for the first phase of the study were processed using a server with 16 GB RAM and two Opteron Dual-Core 2.8 GHz processors at the Center of High Performance Computing (CHPC) at the University of Utah. The CPU processing time to identify pneumonia and influenza cases for both keyword searching and the DCP were 0.21 seconds and 881.83 seconds, respectively. The natural language processing component of the pipeline attributed to 99.4 percent of the DCP processing time (NLP-877 seconds). However, although the processing time for the DCP was about 41 times longer than keyword searching, it still is well within the "in real-time" realm. For instance, in the United States, there are approximately 46,523 deaths; therefore, it would take the DCP 6,364.3 seconds CPU to standardized, code, and flag all the weekly death records.

Statistical Analysis

Keyword Searching

Of the 6,450 records, the keyword searching method flagged 473 records as pneumonia- or influenza-related. From these 473 records, 21 were classified as false positives and 20 as false negatives. Based on these numbers, precision and recall were both calculated at 96%. F-measure was also calculated at 96%. The level of agreement between keyword searching and the reference standard was considered excellent based on the Fleiss' [192] guidelines for characterizing kappas (Cohen's kappa 0.95).

Of the 21 false positives, for 6 of these records, "pneumonia" was present in the cause of death text; however, ICD-10 codes related to pneumonia were absent from their corresponding coded data. Two additional records were flagged because "pneumonia" was included in the substring. The death literals for these two records were "Streptococcal Pneumoniae Septicemia" and "bacteremia due to Streptococcus pneumonia." Data entry errors were attributed to the remaining 13 errors; specifically, all errors were due to "aspiration pneumonia" being classified as a pneumonia-related death. This was due to one of the following: 1) "aspiration" and "pneumonia" being two separate cause of death fields or 2) "pneumonia" not being directly followed by "aspiration" in the death text (example "pneumonia due to secondary aspiration").

For false negatives, this method recorded a total of 20 false negatives. The false negatives generally fell into one of the following two categories: 1) pneumonia being misspelt on the death certificates (n = 8), usually as "pnuemonia" or "pnumonia" and 2) "pneumonia" or "influenza" text being absent from the death phrase, but their corresponding processed code included pneumonia- or influenza-related ICD-10 code(s) (eg J89) (n = 12). F-measure was also calculated at 96%. A high level of agreement was seen among keyword searching and the reference standard (Cohen's kappa 0.95).

Death Certificates Pipeline (DCP)

The Death Certificates Pipeline (DCP) yielded a precision at 98% and recall at 99.8%; false positives and false negatives were 9 and 1, respectively. Like the keyword searching method, 6 of the 9 false positives were due to "pneumonia" being present in one of cause of death fields but not coded into ICD-10. The remaining 3 errors were due to the entry of aspiration pneumonia on the death certificate. The DCP had only 1 false negative for the death literal "recurrent aspiration with pneumonia." F-measure was calculated at 99%. The level of agreement between the pipeline and the reference standard was categorized as excellent because the Cohen's kappa was 0.998.

Statistical analysis of the results showed that two of three standard measures (recall and precision) were all statistically better in the DCP method than keyword searching (Fisher's exact test: recall (*p* = 1.742e-05) and precision (*p* = 0.026). In addition, McNemar's test (*p*-value = 2.152e-05) shows that there is significant difference between the two methods. The performance for each method is given in Table 5.1.

Table 5.1: Comparison of the two methods

Method	Recall	Precision	Cohen's Kappa
Keyword Searching	0.96	0.96	0.96
DCP	0.998	0.98	0.998

Analysis of DCP Failures

Upon analysis of the DCP output, it was observed that most failures were due to discrepancies between the cause of death literal and its respective ICD-10 code. For the reported 9 false positives, 6 were due to "pneumonia" being present in one of cause of death fields but not coded into ICD-10. It can be hypothesized that these 6 false positives were not due to MetaMap or the rules algorithm, but perhaps due to the coding process and the data that provided by UDOH.

As described earlier in the previous chapter, MMDS produces two types of ICD-10 codes: entity axis and record axis codes. From the literature [193], it is suggested that the entity axis codes be used for this type of analysis because they provide the ICD-10 codes for all conditions or events reported on the death certificate [193]; but as mentioned earlier, only the record axis codes were made available for this study. The MMDS applies a rules-based algorithm on the entity axis codes to generate the record axis codes; this produces a better quality of data because conditions are recoded to better match the context of the conditions reported on the death certificate [194]. For example, if "pneumonia with chronic obstructive pulmonary disease" is reported on the death certificate, two conditions will be shown in the entity axis coded data, one reflecting "pneumonia," and the other reflecting "chronic obstructive pulmonary disease." However, the record axis coded data will be replaced with a single condition: "Chronic obstructive pulmonary disease with acute lower respiratory infection" (J44.0). It is important to note that, although we have speculated the reason why the coded data is not included in pneumonia-related codes, we were unable to verify that these codes (codes related to pneumonia) were present in the set of entity axis codes for the six cases.

The final 3 false positives resulted in a second category of errors, the reporting of “aspiration pneumonia” on the death certificate. DCP successfully identified cases where the string “aspiration pneumonia” (C0032290) was reported in the same text field. However, if the strings were reported in separate cause of death fields, the NLP component of the DCP processed the string as two separate texts, thus yielding two codes: one for “aspiration” and the other “pneumonia.” As a result, the record was as a “pneumonia” death; improving the rules of the detection algorithm to be more sophisticated to not flag records like these can overcome this problem.

Results for Phase Two

The results obtained from the feasibility study are summarized and presented in the form of validation through proof-of-concepts by utilizing use case scenarios. These scenarios specify how users carry out their tasks in a specified context and provide examples of usage as an input to design.

Validation Through Proof-of-Concept Studies

Use Case 1: Access and Integrating Public Health Data

In order to illustrate the working principle of the data grid service, we created a scenario that demonstrates the access of public health data that are under a different administrative domain. To simulate a real-world scenario, the grid data services were deployed under different hosting machines: a Linux virtual machine at Center of High Performance Computing and a Windows 7, core i7 machine. To execute this use case, we used 62,181 Utah death records for the years 2003–2007, which were provided by Utah Department of Health. These records were divided by deaths that occurred in counties in the "lower" and "upper" half of the state of Utah, which were then used to populate the two data services. Such distinction resulted in 52,214 death data for the "Nothern Utes" data service, and 9967 for the "Southern Utes" data service. Figure 5.1 shows the division of the state of Utah to provide a visual of the division process. The data services UML models, Figures 5.2 and 5.3, respectively, where both validated in caAdapter, Figure 5.4.

Once the data service was implemented, the user was able to communicate with it using the web interface shown in Figures 5.5 and 5.6. The data service client API was invoked and used to provide an example of querying the data service using the caGrid query language as seen in Figure 5.7.

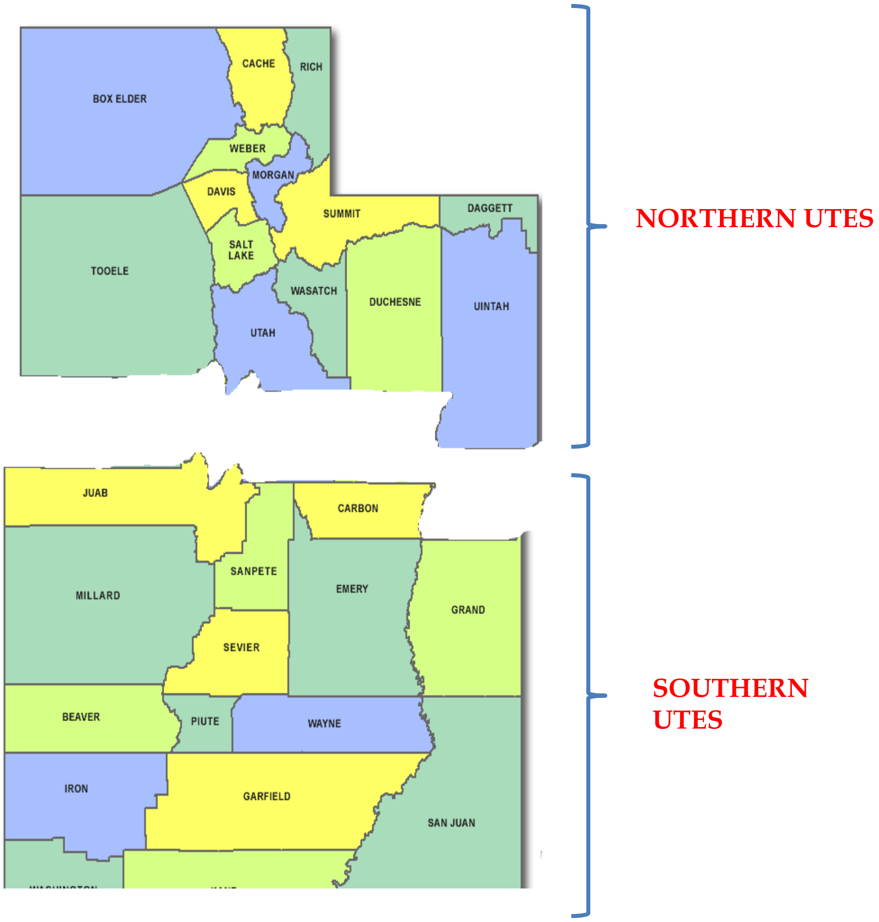


Figure 5.1: "Upper Utah" UML model

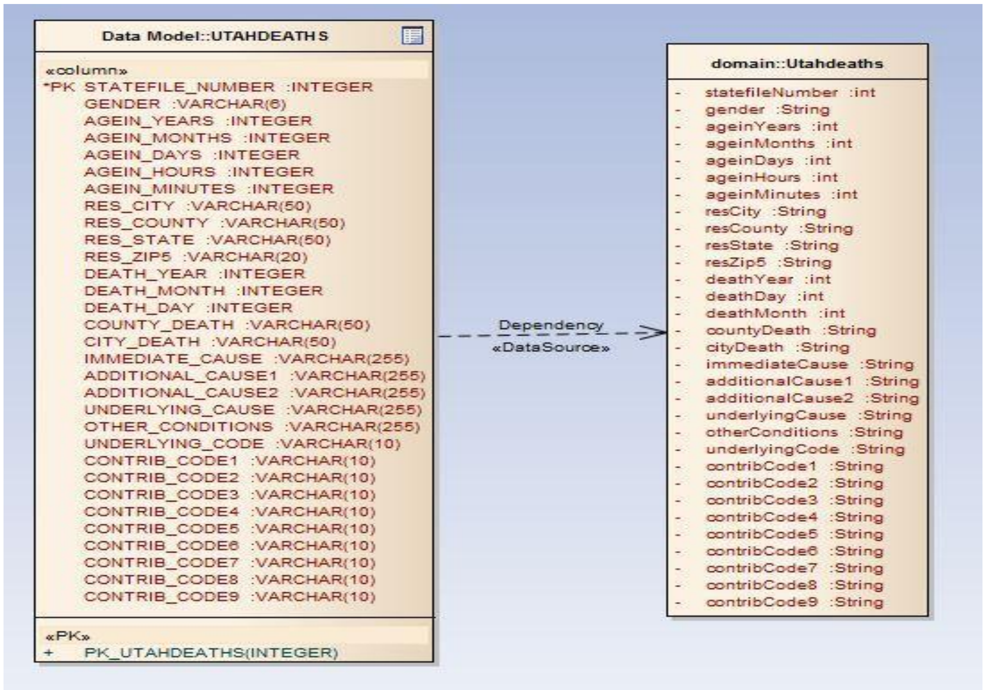


Figure 5.2: “Nothern Utes” UML model

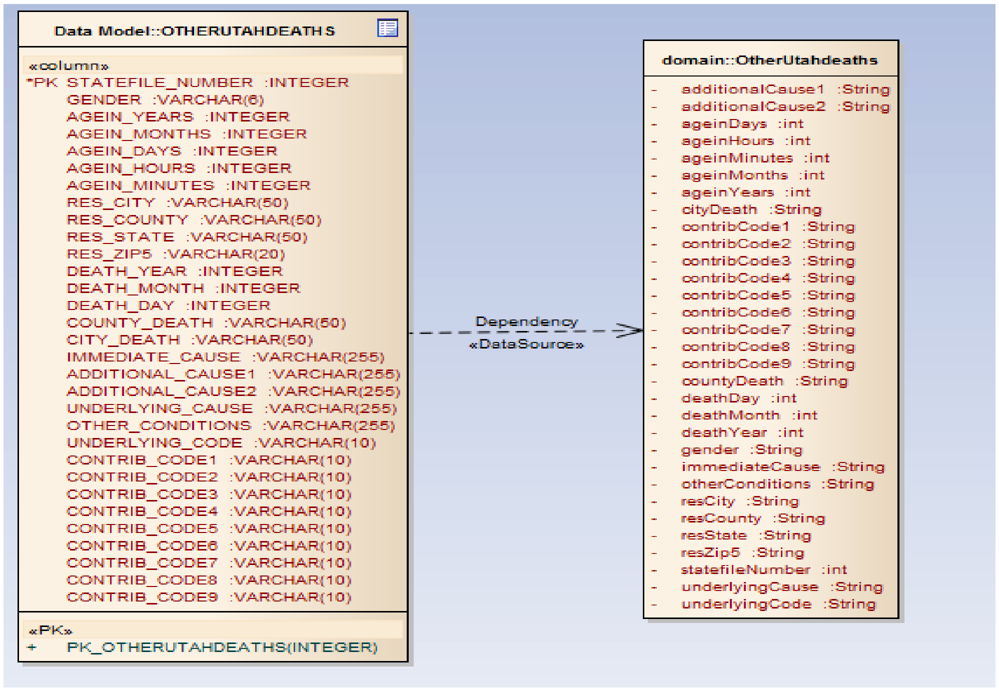


Figure 5.3: “Southern Utes” UML model

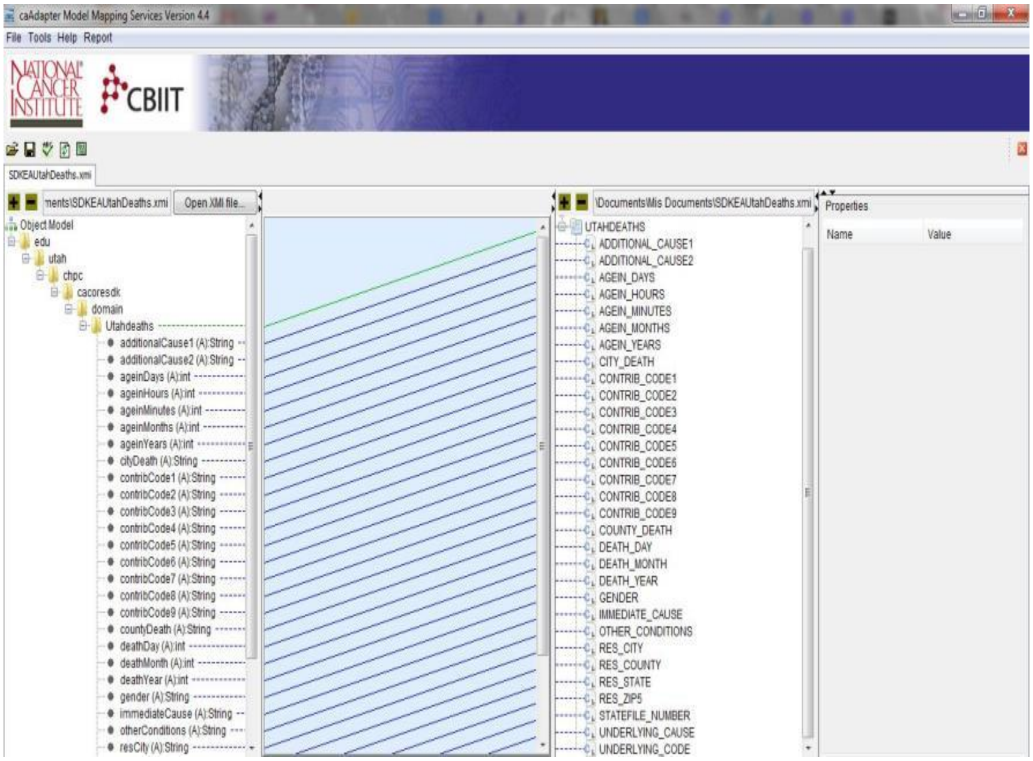
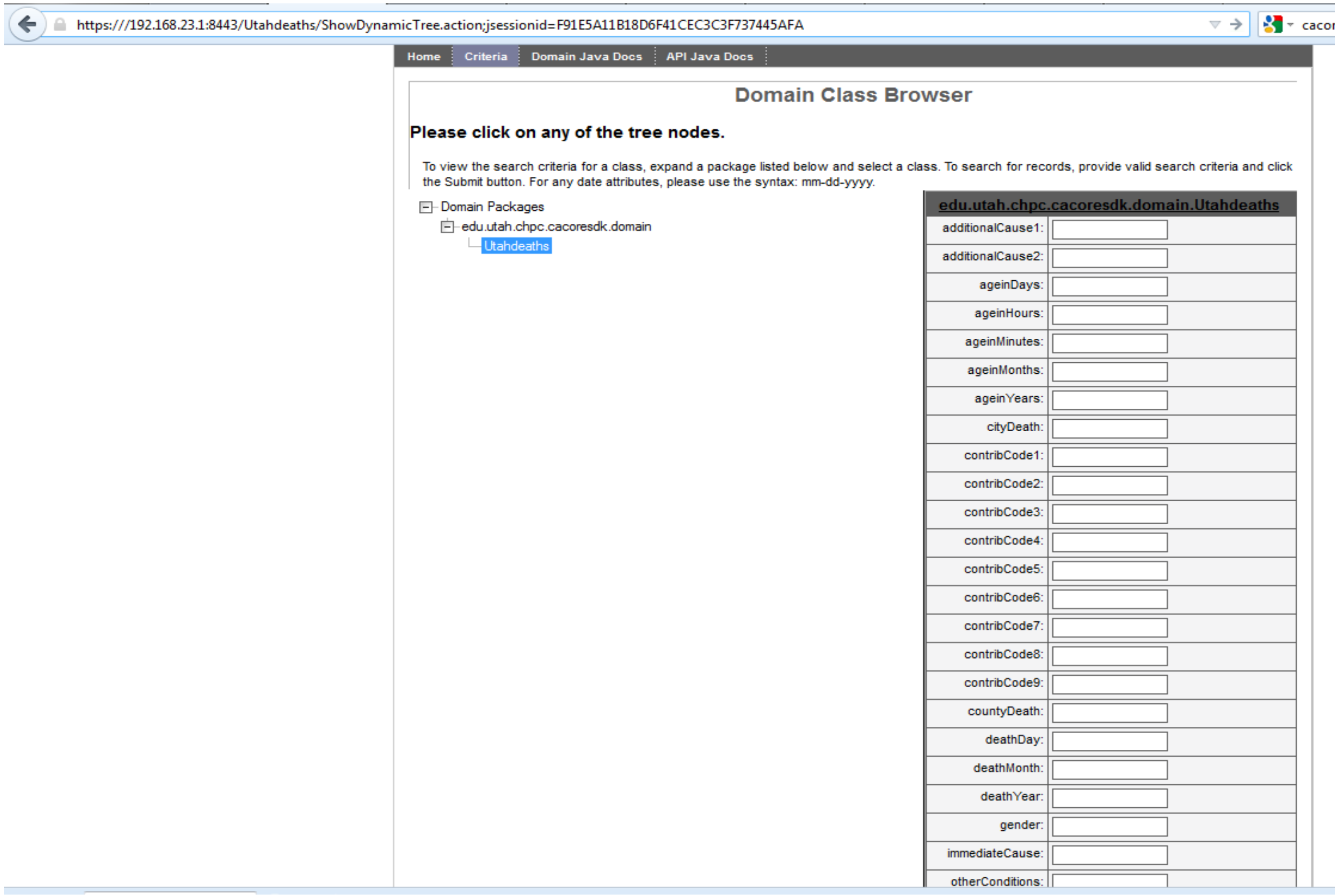


Figure 5.4: Screenshot of validation through caAdapter


https://192.168.23.1:8443/Utahdeaths/ShowDynamicTree.action;jsessionid=F91E5A11B18D6F41CEC3C3F737445AFA
cacor
Home
Criteria
Domain Java Docs
API Java Docs

Domain Class Browser

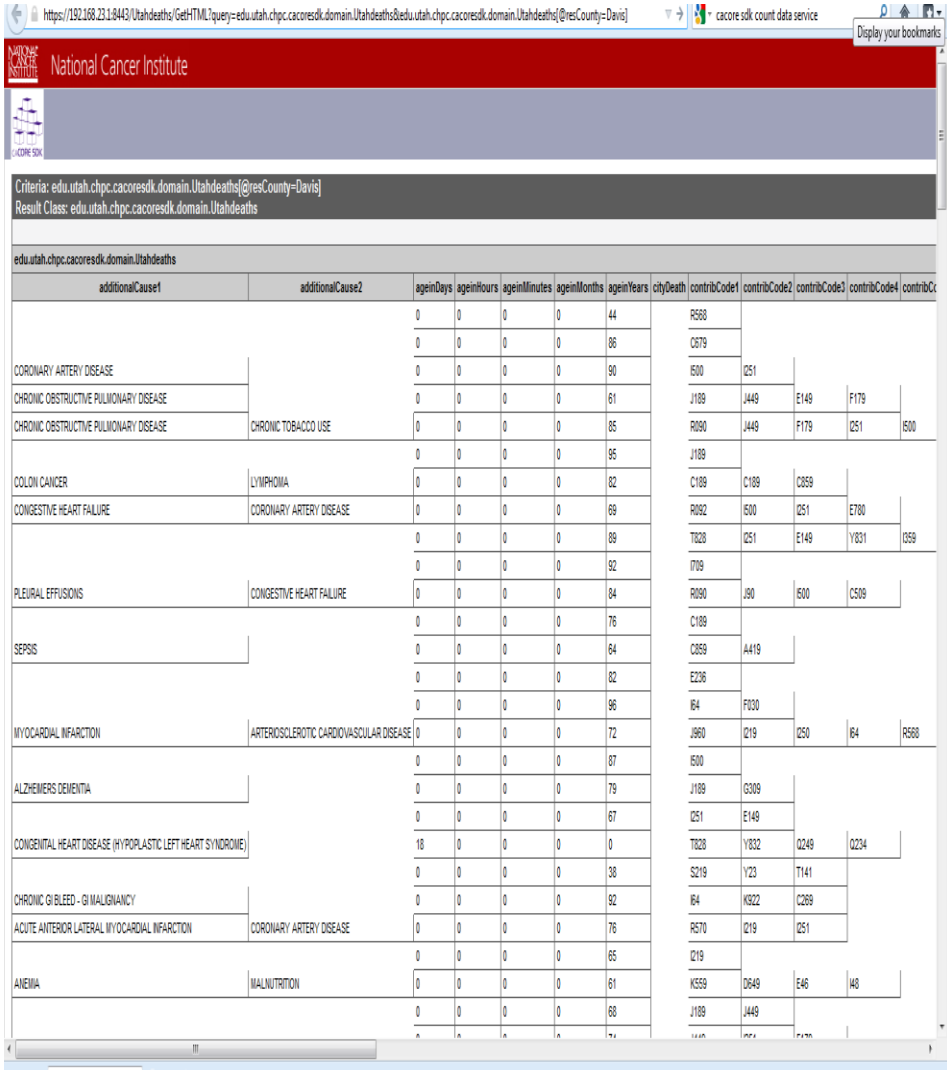
Please click on any of the tree nodes.

To view the search criteria for a class, expand a package listed below and select a class. To search for records, provide valid search criteria and click the Submit button. For any date attributes, please use the syntax: mm-dd-yyyy.

⊟ Domain Packages
⊟ edu.utah.chpc.cacoresdk.domain
Utahdeaths

edu.utah.chpc.cacoresdk.domain.Utahdeaths	
additionalCause1:	
additionalCause2:	
ageinDays:	
ageinHours:	
ageinMinutes:	
ageinMonths:	
ageinYears:	
cityDeath:	
contribCode1:	
contribCode2:	
contribCode3:	
contribCode4:	
contribCode5:	
contribCode6:	
contribCode7:	
contribCode8:	
contribCode9:	
countyDeath:	
deathDay:	
deathMonth:	
deathYear:	
gender:	
immediateCause:	
otherConditions:	

Figure 5.5: caCore SDK web client that provides access to “Upper Salt Lake” data service

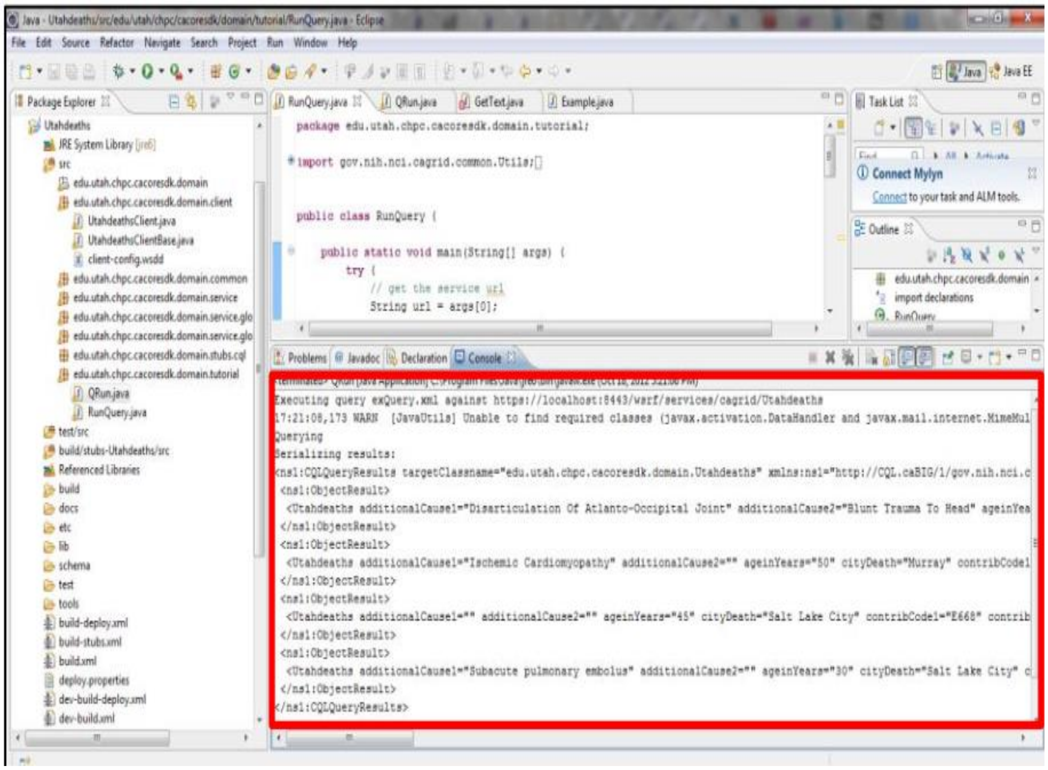

https://192.168.23.1:8443/Utahdeaths/GetHTML?query=edu.utah.chpc.cacoresdk.domain.Utahdeaths&edu.utah.chpc.cacoresdk.domain.Utahdeaths[@resCounty=Davis]
cacore sdk count data service
Display your bookmarks
National Cancer Institute
caCORE SDK

Criteria: edu.utah.chpc.cacoresdk.domain.Utahdeaths[@resCounty=Davis]
Result Class: edu.utah.chpc.cacoresdk.domain.Utahdeaths

edu.utah.chpc.cacoresdk.domain.Utahdeaths

additionalCause1	additionalCause2	ageinDays	ageinHours	ageinMinutes	ageinMonths	ageinYears	cityDeath	contribCode1	contribCode2	contribCode3	contribCode4	contribC
		0	0	0	0	44		R568				
		0	0	0	0	86		C679				
CORONARY ARTERY DISEASE		0	0	0	0	90		I500	I251			
CHRONIC OBSTRUCTIVE PULMONARY DISEASE		0	0	0	0	61		J189	J449	E149	F179	
CHRONIC OBSTRUCTIVE PULMONARY DISEASE	CHRONIC TOBACCO USE	0	0	0	0	85		R090	J449	F179	I251	I500
		0	0	0	0	95		J189				
COLON CANCER	LYMPHOMA	0	0	0	0	82		C189	C189	C859		
CONGESTIVE HEART FAILURE	CORONARY ARTERY DISEASE	0	0	0	0	69		R092	I500	I251	E780	
		0	0	0	0	89		T828	I251	E149	Y831	I359
		0	0	0	0	92		I709				
PLEURAL EFFUSIONS	CONGESTIVE HEART FAILURE	0	0	0	0	84		R090	J90	I500	C509	
		0	0	0	0	76		C189				
SEPSIS		0	0	0	0	64		C859	A419			
		0	0	0	0	82		E236				
		0	0	0	0	96		I64	F030			
MYOCARDIAL INFARCTION	ARTERIOSCLEROTIC CARDIOVASCULAR DISEASE	0	0	0	0	72		J960	I219	I250	I64	R568
		0	0	0	0	87		I500				
ALZHEIMERS DEMENTIA		0	0	0	0	79		J189	G309			
		0	0	0	0	67		I251	E149			
CONGENITAL HEART DISEASE (HYPOPLASTIC LEFT HEART SYNDROME)		18	0	0	0	0		T828	Y832	Q249	Q234	
		0	0	0	0	38		S219	Y23	T141		
CHRONIC GI BLEED - GI MALIGNANCY		0	0	0	0	92		I64	K922	C269		
ACUTE ANTERIOR LATERAL MYOCARDIAL INFARCTION	CORONARY ARTERY DISEASE	0	0	0	0	76		R570	I219	I251		
		0	0	0	0	65		I219				
ANEMIA	MALNUTRITION	0	0	0	0	61		K559	D649	E46	I48	
		0	0	0	0	68		J189	J449			

Figure 5.6: Querying data service through web-interface



Java - Utahdeaths/src/edu/utah/chpc/cacoresdk/domain/tutorial/RunQuery.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Java
Java EE
Package Explorer
Utahdeaths
JRE System Library [jre6]
src
edu.utah.chpc.cacoresdk.domain
edu.utah.chpc.cacoresdk.domain.client
UtahdeathsClient.java
UtahdeathsClientBase.java
client-config.wsdd
edu.utah.chpc.cacoresdk.domain.common
edu.utah.chpc.cacoresdk.domain.service
edu.utah.chpc.cacoresdk.domain.service.glo
edu.utah.chpc.cacoresdk.domain.service.glo
edu.utah.chpc.cacoresdk.domain.stubs.cql
edu.utah.chpc.cacoresdk.domain.tutorial
QRun.java
RunQuery.java
test/src
build/stubs-Utahdeaths/src
Referenced Libraries
build
docs
etc
lib
schema
test
tools
build-deploy.xml
build-stubs.xml
build.xml
deploy.properties
dev-build-deploy.xml
dev-build.xml
RunQuery.java
QRun.java
GetText.java
Example.java
Task List
Find
All
Activate
Connect Mylyn
Connect to your task and ALM tools.
Outline
edu.utah.chpc.cacoresdk.domain
import declarations
RunQuery
Problems
Javadoc
Declaration
Console

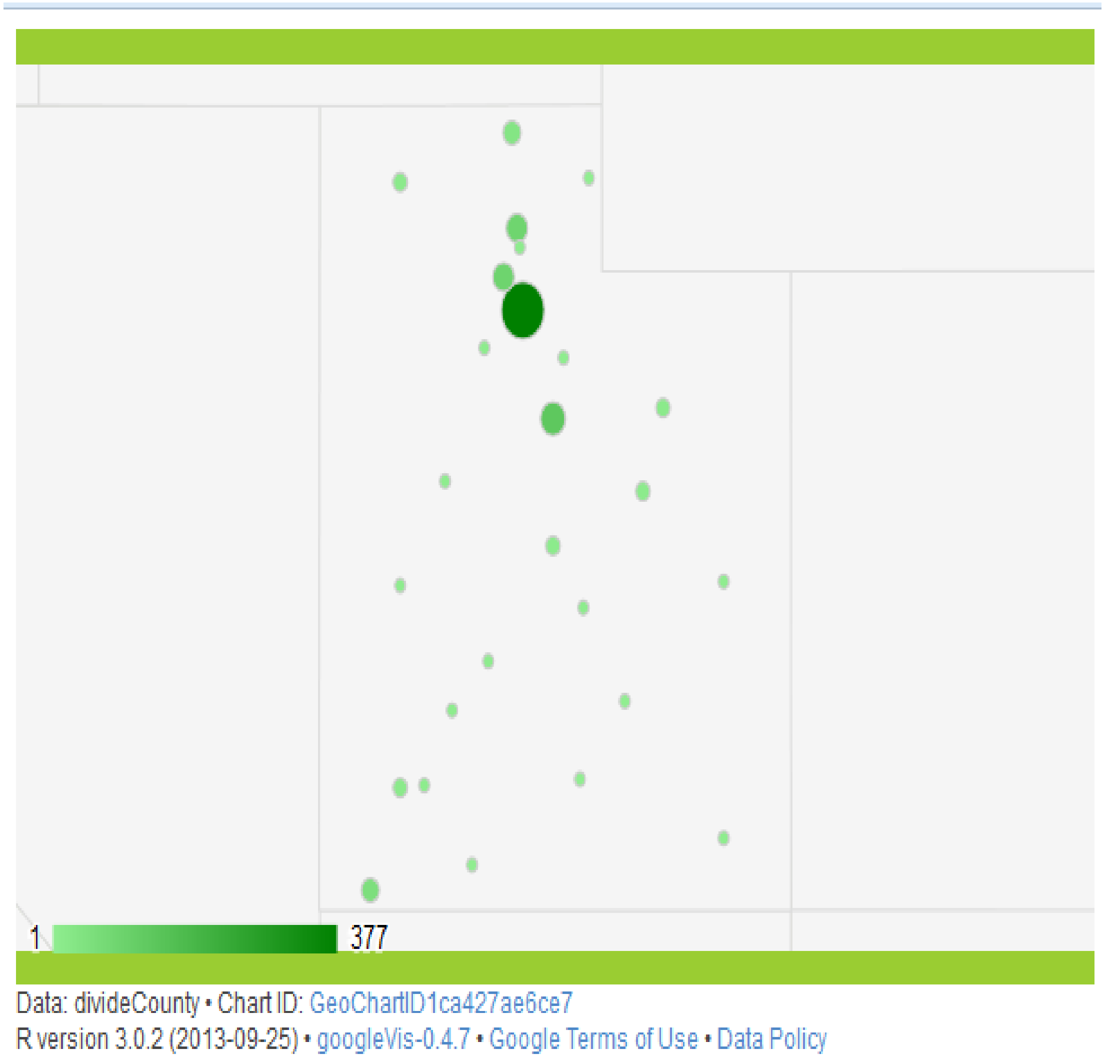
```
package edu.utah.chpc.cacoresdk.domain.tutorial;

import gov.nih.nci.cagrid.common.Utils;


public class RunQuery {

    public static void main(String[] args) {
        try {
            // get the service url
            String url = args[0];
```

```
<terminated> QRun [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Oct 18, 2012 5:21:06 PM)
Executing query exQuery.xml against https://localhost:8443/wsrf/services/cagrid/Utahdeaths
17:21:08,173 WARN  [JavaUtils] Unable to find required classes (javax.activation.DataHandler and javax.mail.internet.MimeMul
Querying
Serializing results:
<ns1:CQLQueryResults targetClassname="edu.utah.chpc.cacoresdk.domain.Utahdeaths" xmlns:ns1="http://CQL.caBIG/1/gov.nih.nci.c
 <ns1:ObjectResult>
  <Utahdeaths additionalCause1="Disarticulation Of Atlanto-Occipital Joint" additionalCause2="Blunt Trauma To Head" ageinYea
 </ns1:ObjectResult>
 <ns1:ObjectResult>
  <Utahdeaths additionalCause1="Ischemic Cardiomyopathy" additionalCause2="" ageinYears="50" cityDeath="Murray" contribCode1
 </ns1:ObjectResult>
 <ns1:ObjectResult>
  <Utahdeaths additionalCause1="" additionalCause2="" ageinYears="45" cityDeath="Salt Lake City" contribCode1="E668" contrib
 </ns1:ObjectResult>
 <ns1:ObjectResult>
  <Utahdeaths additionalCause1="Subacute pulmonary embolus" additionalCause2="" ageinYears="30" cityDeath="Salt Lake City" c
 </ns1:ObjectResult>
</ns1:CQLQueryResults>
```

Figure 5.7: Querying data service using DCQL

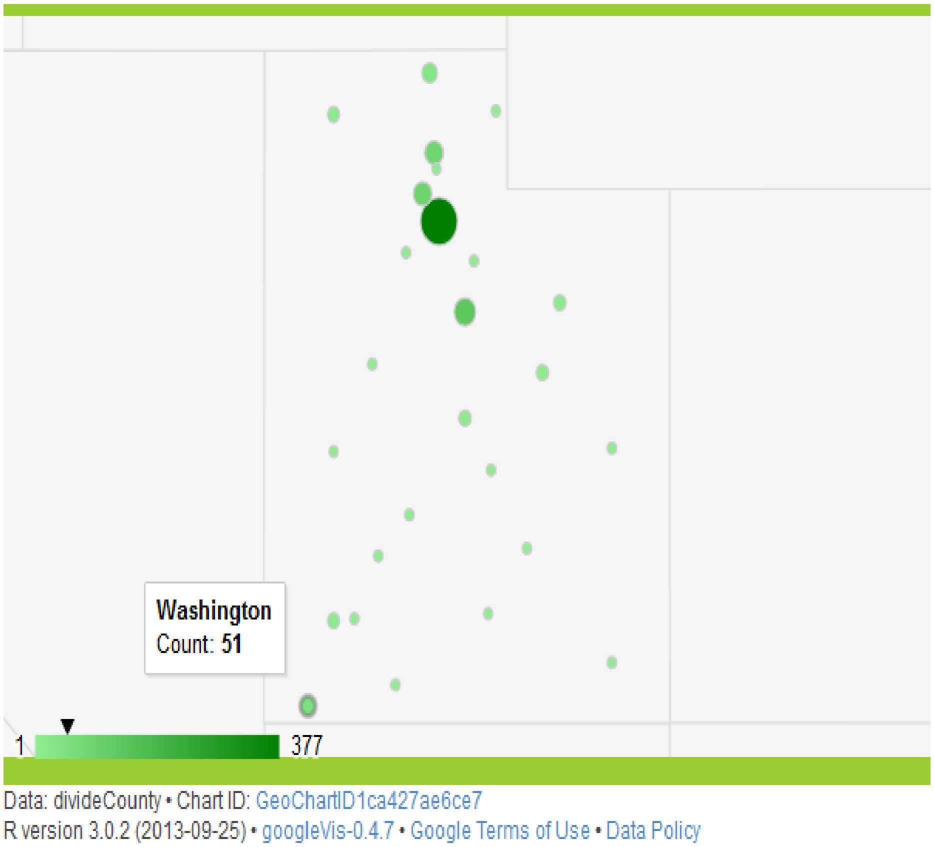
The purpose of this use case is to demonstrate that the data service can be queried and the results can be analyzed in an analytical tool of choice, in this case, R. Therefore, the queried results were imported into R and the R package "GoogleVis" was used to the plot the total counts by counties (Figure 5.8).

The Google Chart Tools offer interactive charts that can be embedded into web pages; this is depicted in Figure 5.9, where the user hovered over Washington County to view the total number of deaths for this specific county.



Data: divideCounty • Chart ID: GeoChartID1ca427ae6ce7
R version 3.0.2 (2013-09-25) • googleVis-0.4.7 • Google Terms of Use • Data Policy

Figure 5.8: Plotting total count by county using R



Data: divideCounty • Chart ID: GeoChartID1ca427ae6ce7
R version 3.0.2 (2013-09-25) • googleVis-0.4.7 • Google Terms of Use • Data Policy

Figure 5.9: Example of “GoogleVis” interactive plot

Use Case 2: Access to Natural Language Processing Tools

The MetaMap grid version was invoked and deployed to the caGrid training grid. This analytical grid service can be primarily used in three interaction modes: programmatically, web interface, and/or scientific workflows. A brief description of each interactive mode follows:

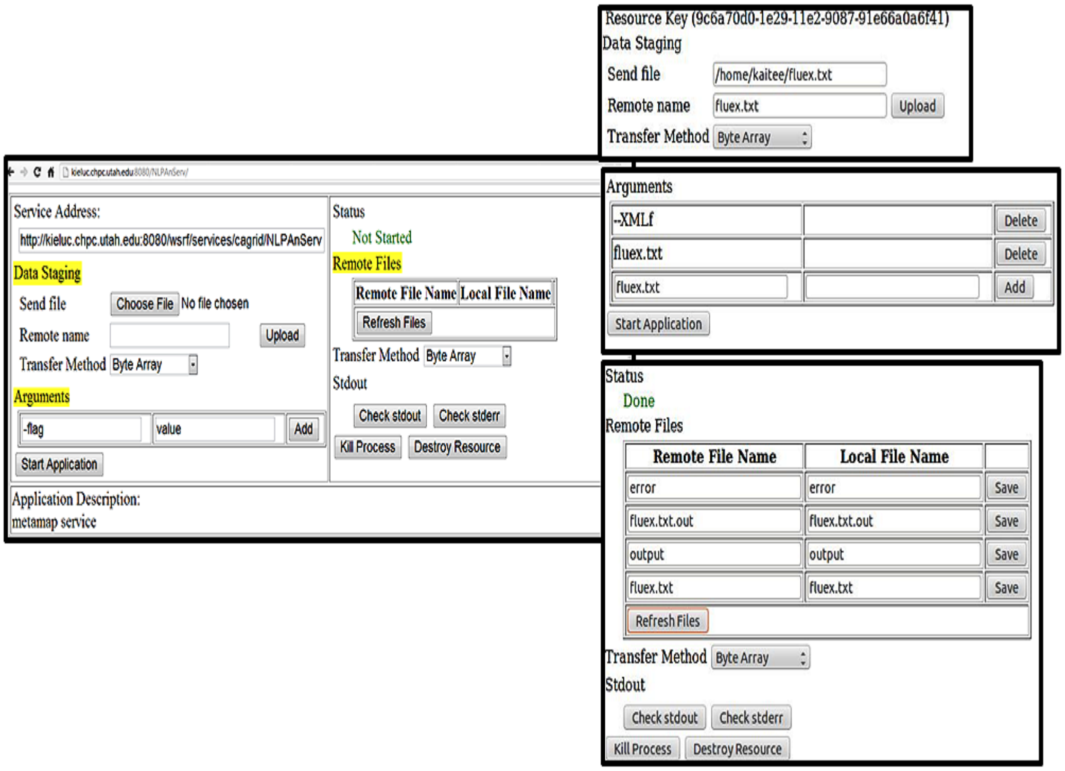
1. Programmatically: Through the MetaMap grid application interface (API), developers have greater flexibility by using programming languages or scripts to perform public health-related analysis.
2. Web Interface: The web interface is discussed in the succeeding section.
3. Workflows: Like many scientific domains, public health analyses are becoming

> increasingly data and process driven; this type of analysis is well modeled by scientific workflows, scientific workflows were explained in Chapter 4. Consider the example of mortality surveillance that is constantly mentioned throughout this dissertation; a workflow can be created to automatically standardize and code death certificates and include operations to identify diseases of interest.

Natural language processing tools, like MetaMap, are very technical my nature. They usually require users to have technical, programmatic, and/or command line knowledge to the run these types of software. As a result, for the purpose of this study, our validation was done through the web interface that was automatically generated by gRavi; the web interface can be accessed at http://kieluc.chpc.utah.edu:8080/NLPAnServ. The gRavi web client has three distinct sections: 1) data staging: used to upload files required to the working directory on the execution machine; 2) remote files: used to view all files in the working directory on the execution machine, and 3) arguments section: used to add MetaMap command-line arguments. This web interface was used to execute the scenario described below.

Use case scenario. An epidemiologist wants to process unstructured, free text death certificates to assess the severity of a pneumonia outbreak. On the client's local machine, he/she has available software/algorithms that will be used to identify the disease of interest. However, due to limitations, such as lack of computational resources to quickly process the death certificates, the client is unable to properly analyze the unstructured data. This client can use the secure web service to upload the death certificates to execute the text processing remotely, and transfer the results back to the client.

To execute the scenario above, the client web interface was used because we thought it would be the best option for demonstrating the service’s usefulness to public health officials because, of the three options mentioned above, the web client provides a greater ease of use for the target user over programmatic interactions. Figure 5.10 shows the web client servlet; the three distinct sections mentioned above are highlighted on the right.



kieluc.chpc.utah.edu:8080/NLPAnServ/

Service Address: http://kieluc.chpc.utah.edu:8080/wsrf/services/cagrid/NLPAnServ

Data Staging

Send file: Choose File No file chosen

Remote name: Upload

Transfer Method: Byte Array

Arguments

-flag | value | Add

Start Application

Status: Not Started

Remote Files

Remote File Name	Local File Name

Refresh Files

Transfer Method: Byte Array

Stdout

Check stdout | Check stderr

Kill Process | Destroy Resource

Application Description: metamap service

Resource Key (9c6a70d0-1e29-11e2-9087-91e66a0a6f41)

Data Staging

Send file: /home/kaitee/fluex.txt

Remote name: fluex.txt Upload

Transfer Method: Byte Array

Arguments

--XMLf		Delete
fluex.txt		Delete
fluex.txt		Add

Start Application

Status: Done

Remote Files

Remote File Name	Local File Name	
error	error	Save
fluex.txt.out	fluex.txt.out	Save
output	output	Save
fluex.txt	fluex.txt	Save

Refresh Files

Transfer Method: Byte Array

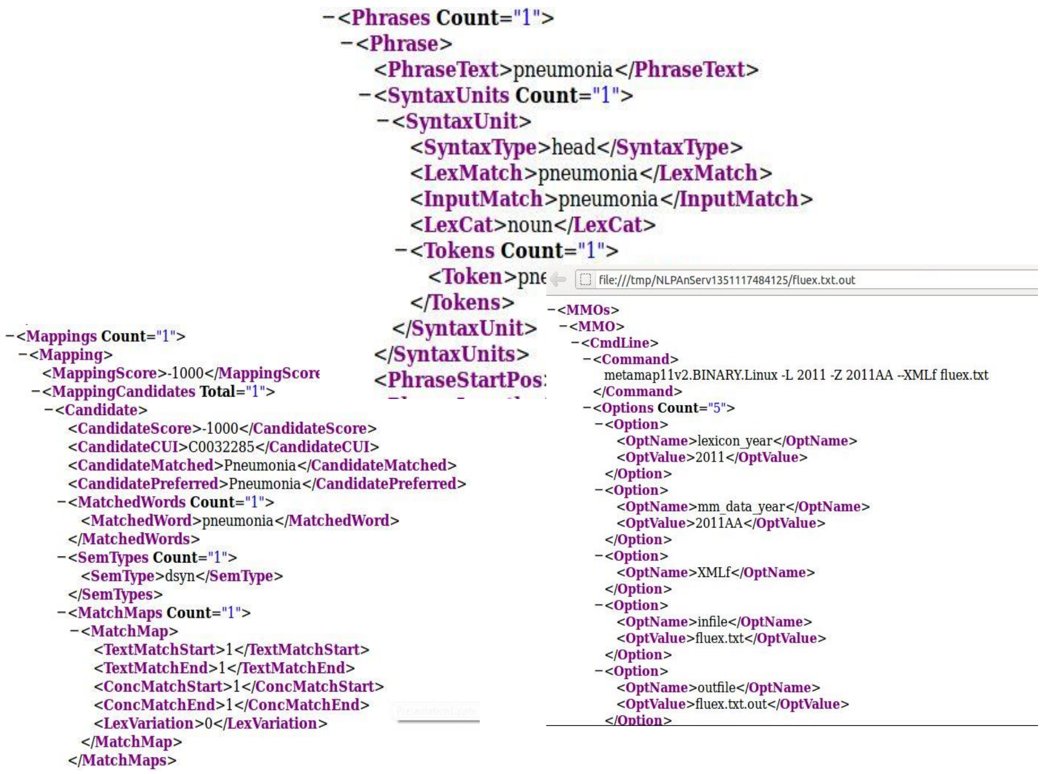
Stdout

Check stdout | Check stderr

Kill Process | Destroy Resource

Figure 5.10: Snapshot of processing a file with MetaMap grid web client. Adapted from: Davis, Kailah, Ronald C. Price, and Julio C. Facelli. "Implementing public health analytical services: Grid enabling of MetaMap." In Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on, pp. 113–118. IEEE, 2013. © 2013 IEEE. Reprinted with permission.

For the user scenario, phase one study data was used to evaluate the performance of the grid enabled MetaMap. Using the grid web client, we uploaded a text file of the death phrases for processing. The application then ran the text file on the execution server; this machine had all the necessary MetaMap services running to process the file and execute the specific scenario. Result of the analysis was returned to the user in the form of electronic files; in this case the fluextxt.out file. A snippet of the output can be seen in Figure 5.11.



```
−<Phrases Count="1">
  −<Phrase>
     <PhraseText>pneumonia</PhraseText>
    −<SyntaxUnits Count="1">
      −<SyntaxUnit>
         <SyntaxType>head</SyntaxType>
         <LexMatch>pneumonia</LexMatch>
         <InputMatch>pneumonia</InputMatch>
         <LexCat>noun</LexCat>
        −<Tokens Count="1">
           <Token>pne
         </Tokens>
       </SyntaxUnit>
     </SyntaxUnits>
     <PhraseStartPos
```

```
−<Mappings Count="1">
  −<Mapping>
     <MappingScore>-1000</MappingScore
    −<MappingCandidates Total="1">
      −<Candidate>
         <CandidateScore>-1000</CandidateScore>
         <CandidateCUI>C0032285</CandidateCUI>
         <CandidateMatched>Pneumonia</CandidateMatched>
         <CandidatePreferred>Pneumonia</CandidatePreferred>
        −<MatchedWords Count="1">
           <MatchedWord>pneumonia</MatchedWord>
         </MatchedWords>
        −<SemTypes Count="1">
           <SemType>dsyn</SemType>
         </SemTypes>
        −<MatchMaps Count="1">
          −<MatchMap>
             <TextMatchStart>1</TextMatchStart>
             <TextMatchEnd>1</TextMatchEnd>
             <ConcMatchStart>1</ConcMatchStart>
             <ConcMatchEnd>1</ConcMatchEnd>
             <LexVariation>0</LexVariation>
           </MatchMap>
         </MatchMaps>
```

file:///tmp/NLPAnServ1351117484125/fluex.txt.out

```
−<MMOs>
  −<MMO>
    −<CmdLine>
      −<Command>
         metamap11v2.BINARY.Linux -L 2011 -Z 2011AA --XMLf fluex.txt
       </Command>
      −<Options Count="5">
        −<Option>
           <OptName>lexicon_year</OptName>
           <OptValue>2011</OptValue>
         </Option>
        −<Option>
           <OptName>mm_data_year</OptName>
           <OptValue>2011AA</OptValue>
         </Option>
        −<Option>
           <OptName>XMLf</OptName>
         </Option>
        −<Option>
           <OptName>infile</OptName>
           <OptValue>fluex.txt</OptValue>
         </Option>
        −<Option>
           <OptName>outfile</OptName>
           <OptValue>fluex.txt.out</OptValue>
         </Option>
```

Figure 5.11: Snippet of returned file. Adapted from: Davis, Kailah, Ronald C. Price, and Julio C. Facelli. "Implementing public health analytical services: Grid enabling of MetaMap." In Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on, pp. 113–118. IEEE, 2013. © 2013 IEEE. Reprinted with permission.

Using the grid version of MetaMap gave exactly the same results as those executed in the stand-alone version. Moreover, the difference between execution times on the local installation (~0.169 seconds per death literals) and the grid version of MetaMap (0.132 seconds per death literal) were negligible. For this project, we did not perform a rigorous analysis of the time and effort needed to implement MetaMap in a grid environment.

The results of from this section were originally published in the paper by Davis et al. [190] © 2013 IEEE.

Use Case 3: Deploy Commonly Used Public Health Surveillance Algorithms in a Grid Environment

The overall aim is to develop an analytical service that allows users to submit R functions for execution in a grid environment. Creating a grid version of R allows users such as epidemiologists, and others unfamiliar with the statistical language, to perform powerful statistical analyses easily. Also, it demonstrates the functionality of utilizing a grid to provide access to tools that may be used at one department but not another. This example, along with the second aims, demonstrates the feasibility of utilizing a grid to provide collaborative model/architecture to help increase collaboration within public health environs.

Use case scenario. An epidemiologist wants to perform aberration detection in an R environment. On his local machine (client), he uses the R interface to create a script with his choice of surveillance algorithm, which he wishes to use on a public health data set. Although the client could analyze the data on his local machine, there might be instances where the data are too large and execution in a grid environment might have bigger computation power than the client's machines, thus making it more efficient to execute the algorithm remotely and transfer the results to the client.

The implementation of the grid-based version of R (rGRID) followed the methods for creating the MetaMap grid and produced the three distinct user interaction modes: a user interface, workflows, and programmatically. Therefore, in-depth details about these steps will not be provided. However, it is important to note that users of the R grid environment will have access to the distributed statistical in a transparent fashion, as if the software was local. The complexity of the grid is thus hidden from the user. In practice, grid access can be performed by passing the R code or script to the grid as an input along with the data to be processed, if necessary; during this process, a working directory is created on the remote machine where the R script and data are transferred to (to a GridFTP server later used for file staging). The job is then submitted to the grid system and after the execution is finished, the resulting output files (staged out by the grid system) are transferred to the client; the results are written with an "**out"** at the end of each file. It is important to note that while the remote execution is active, and the user waits for the results, the create output **(out)** file is locked. The different execution steps are shown in Figure 5.12.

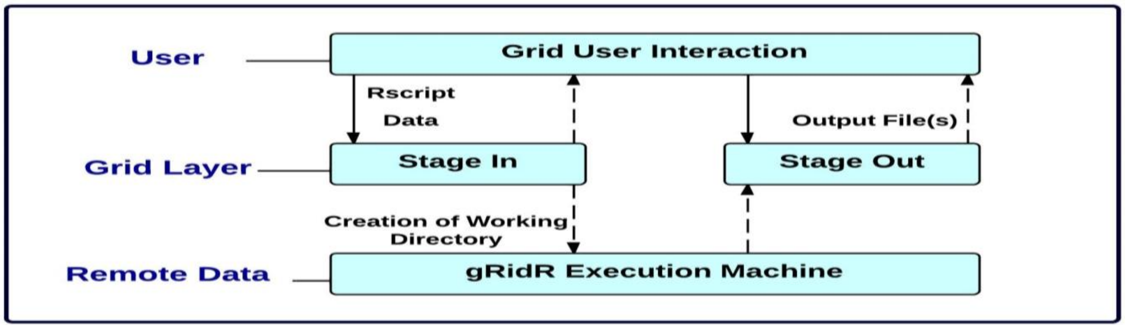


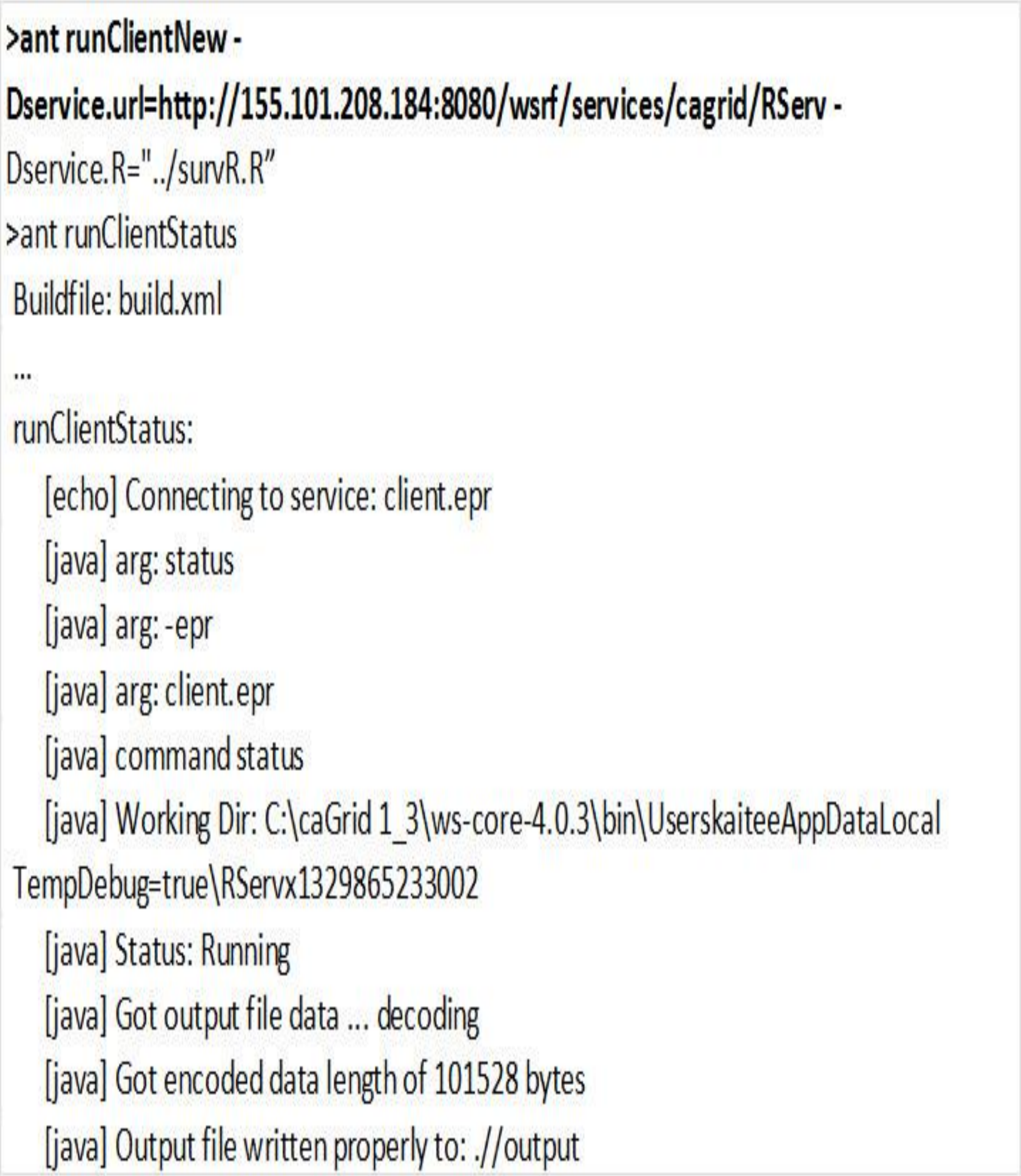
Figure 5.12: Execution of grid-based version of R

Scenario

To demonstrate the working principles of the R in a grid environment, in an article by Hohle et al., a simple example of aberration detection in mortality data and for which all data are available online [195]. This also provides a validation of the gird-based version of R in a realistic usage scenario.

In the "mortality scenario," Danish 1994–2008 mortality data containing the weekly number of all-cause mortality are used to demonstrate the plot functions and different aberration algorithms in the "surveillance" package. The Danish mortality data were produced from the European monitoring of excess mortality for public health action (EuroMOMO) project; more information about the EuroMOMO project can be found in the selected article (Hohle et al. [195]). The work will be validated by showing that the results from Hohle et al. can be reproduced using the grid-based version of R (rGRID).

To execute this use case, we used Globus to send an R script to the execution machine. In the script, we loaded the "surveillance" package, performed a subset of the analyses found in the literature (Hohle et al.), CUSUM and Farrington algorithms, and created a "pdf" file with all the necessary graphs; this file was then transferred back to the user. The execution command can be seen in Figure 5.13. After executing the command, the rGRID generated graphs for the Farrington and CUSUM algorithms, Figures 5.14 and 5.15, that were identical to those in the literature, Hohle et al. [195], thus validating the implementation, particularly for implementing the "surveillance" package that implements analysis steps for identifying outbreaks in mortality data. Moreover, this proof-of-concept showed that other useful R commands such as export graphs to a "pdf" file can be successfully executed in rGRID.



```
>ant runClientNew -
Dservice.url=http://155.101.208.184:8080/wsrf/services/cagrid/RServ -
Dservice.R="../survR.R"
>ant runClientStatus
 Buildfile: build.xml
 ...
 runClientStatus:
     [echo] Connecting to service: client.epr
     [java] arg: status
     [java] arg: -epr
     [java] arg: client.epr
     [java] command status
     [java] Working Dir: C:\caGrid 1_3\ws-core-4.0.3\bin\UserskaiteeAppDataLocal
 TempDebug=true\RServx1329865233002
     [java] Status: Running
     [java] Got output file data ... decoding
     [java] Got encoded data length of 101528 bytes
     [java] Output file written properly to: .//output
```

Figure 5.13: Command used to run the R Script on the grid-based version of R

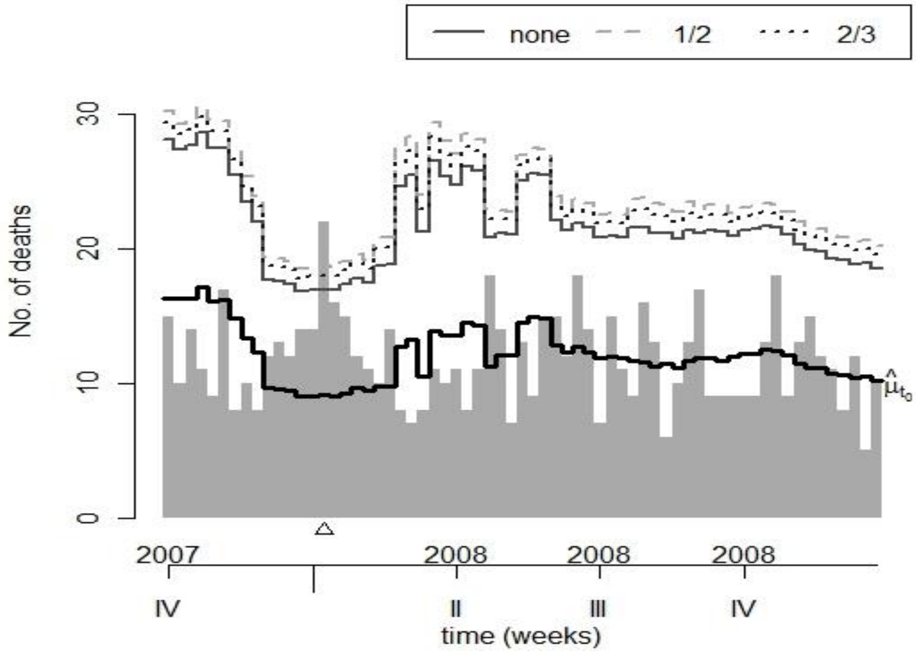


Figure 5.14: Farrington graph produced by rGRID

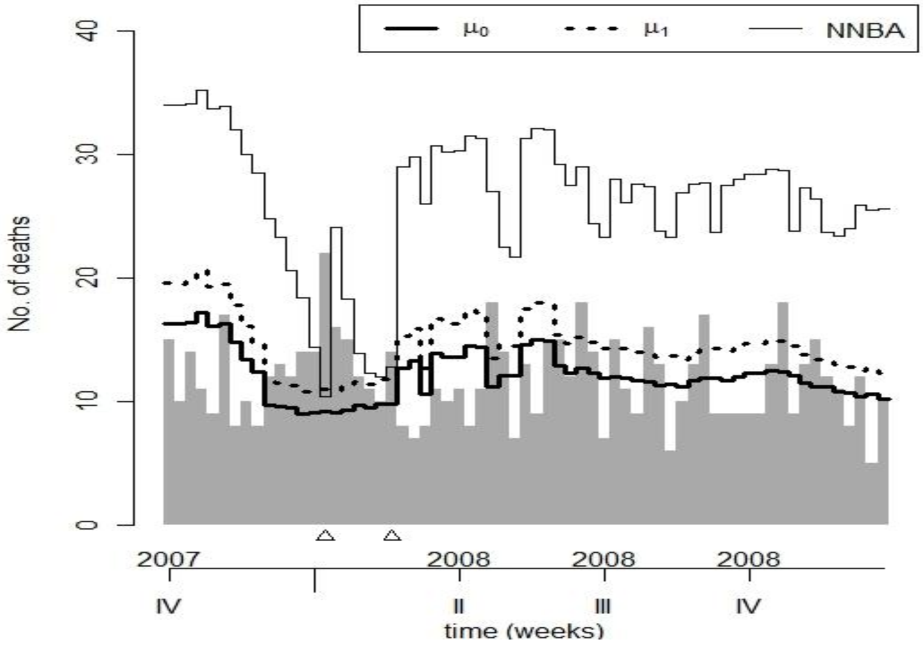
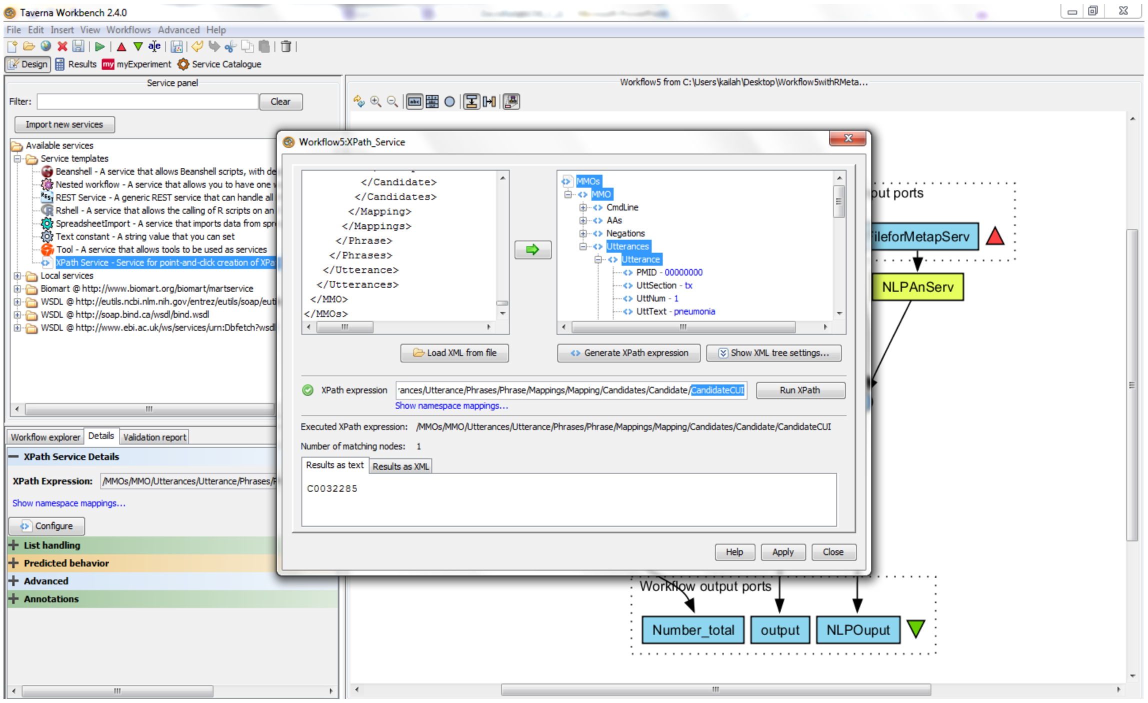


Figure 5.15: CUSUM graph produced by rGRID

Use Case 4: Replicate Current Mortality Surveillance Workflow

As noted in the previous chapter, a Taverna workflow was constructed to model the CDC's 122 Cities Mortality Reporting System pipeline and was tested using death data provided by the Utah Department of Health. For this workflow, the input is an XML document a user got from the "NLPAnServ" grid service; the input is then passed to the "XPath_Service," which parses the XML and identify codes from the "Candidate -> Mapping Tree"; this step can be seen in Figure 5.16.

Once the relevant codes are identified, the data are sent to a nested workflow; in this nested workflow, the list of codes is read into as a vector and fed into the "Rshell." The "Rshell" contains the necessary R code to identify the total number for codes of interest. The count is then output to a MySQL database (Figure 5.17), as well as a text file on the user's computer (Figure 5.18).



Filter:

Available services
- Service templates
  - Beanshell - A service that allows Beanshell scripts, with de
  - Nested workflow - A service that allows you to have one
  - REST Service - A generic REST service that can handle all
  - Rshell - A service that allows the calling of R scripts on an
  - SpreadsheetImport - A service that imports data from spr
  - Text constant - A string value that you can set
  - Tool - A service that allows tools to be used as services
  - XPath Service - Service for point-and-click creation of XPa
- Local services
- Biomart @ http://www.biomart.org/biomart/martservice
- WSDL @ http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/euti
- WSDL @ http://soap.bind.ca/wsdl/bind.wsdl
- WSDL @ http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl

XPath Service Details

XPath Expression: /MMOs/MMO/Utterances/Utterance/Phrases/P

```
          </Candidate>
         </Candidates>
        </Mapping>
       </Mappings>
      </Phrase>
     </Phrases>
    </Utterance>
   </Utterances>
  </MMO>
</MMOs>
```

```
MMOs
  MMO
    CmdLine
    AAs
    Negations
    Utterances
      Utterance
        PMID - 00000000
        UttSection - tx
        UttNum - 1
        UttText - pneumonia
```

XPath expression: ·ances/Utterance/Phrases/Phrase/Mappings/Mapping/Candidates/Candidate/CandidateCUI

Executed XPath expression: /MMOs/MMO/Utterances/Utterance/Phrases/Phrase/Mappings/Mapping/Candidates/Candidate/CandidateCUI

Number of matching nodes: 1

```
C0032285
```

Figure 5.16: Example parsing of XML to find "MetaMapping" CUI

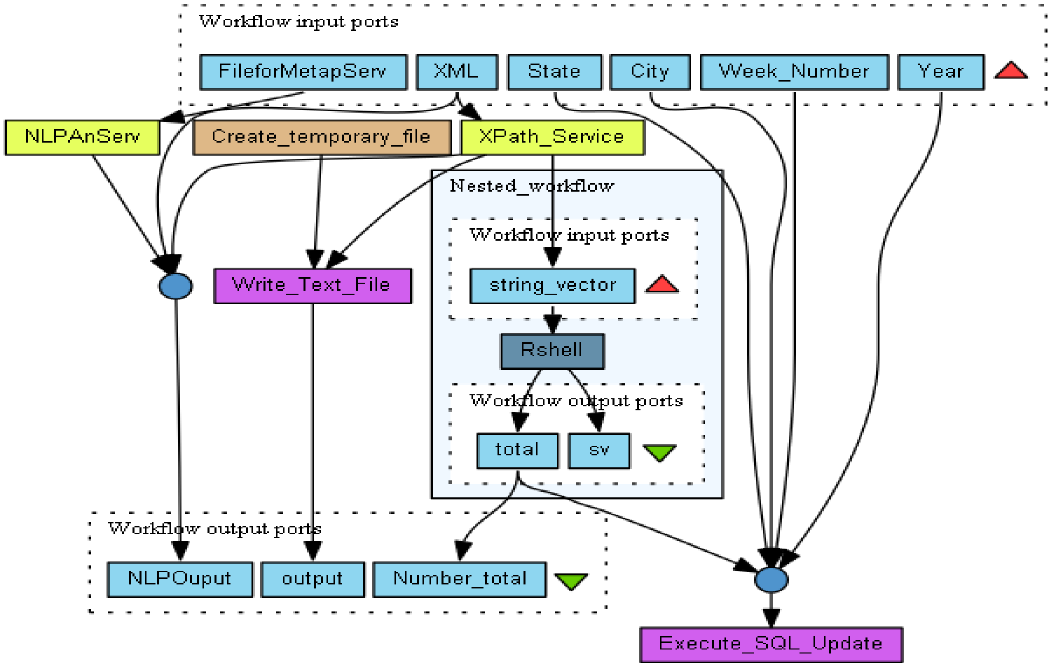


Figure 5.17: Taverna workflow output to MySQL Database

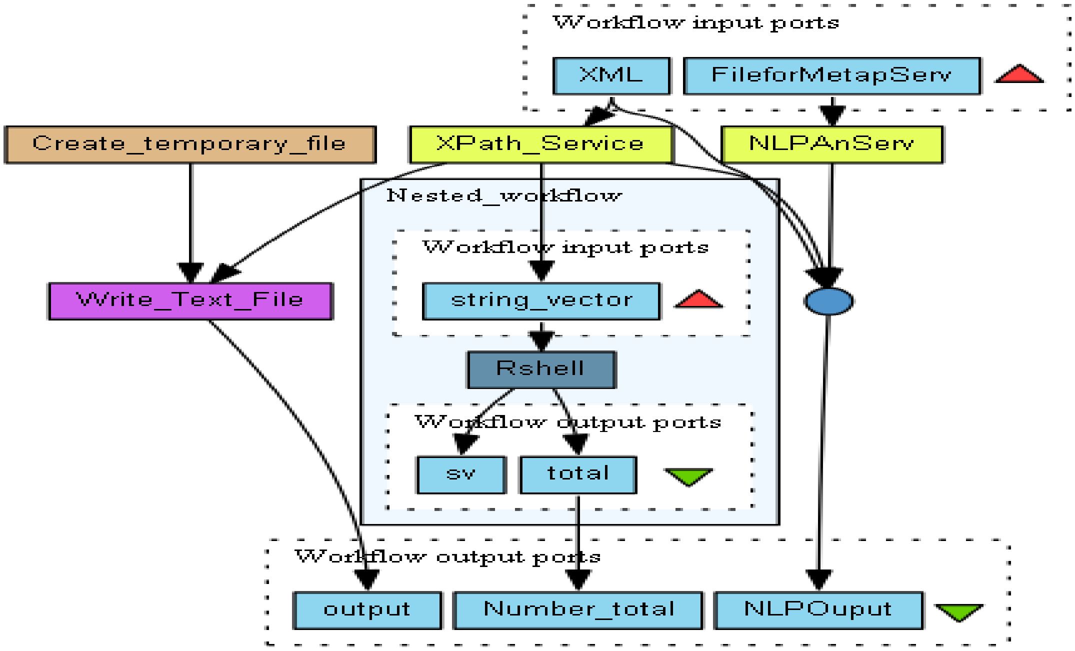
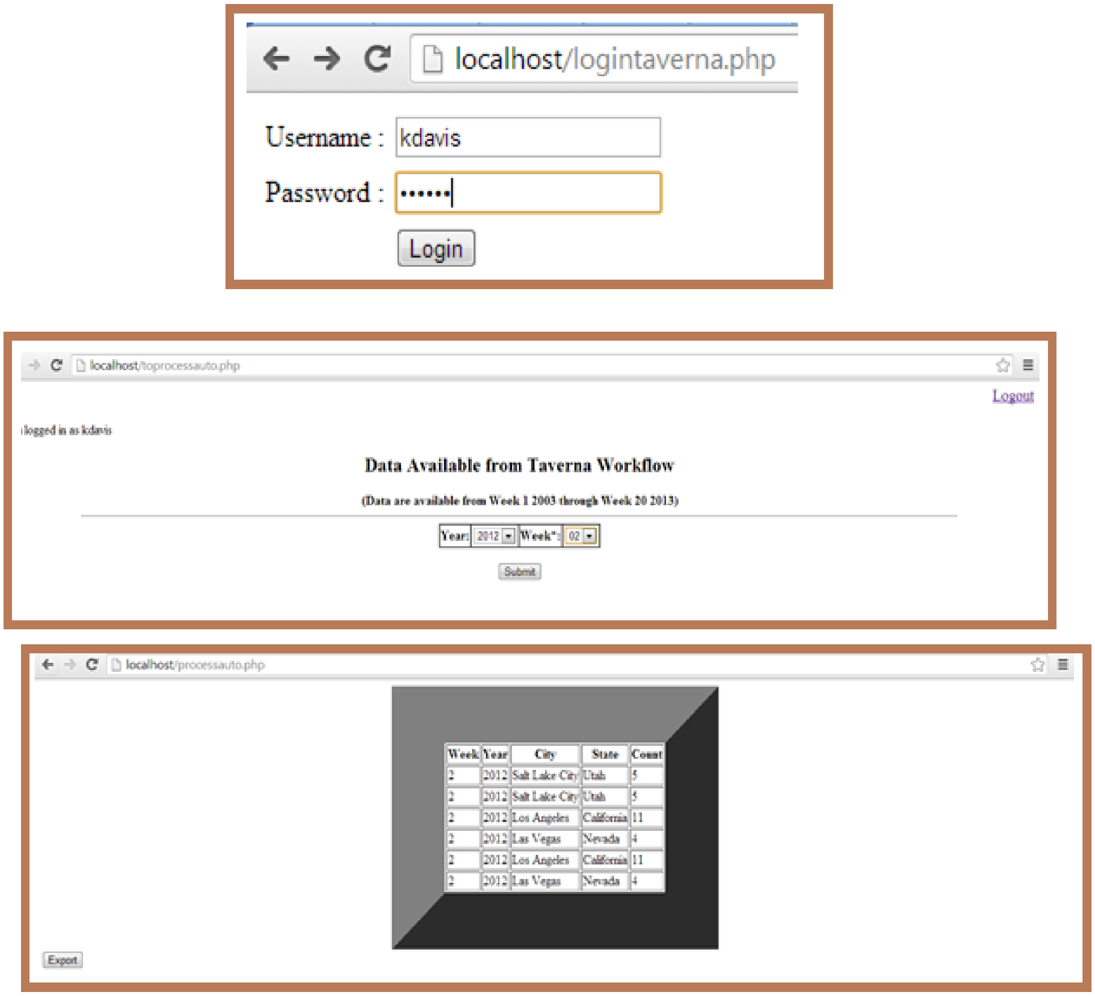


Figure 5.18: Taverna workflow output to text file

When outputting to the database, a SQL update query is updated with the "*total count*," "*State*," "*City*," "*Week Number*," and "*Year*." These data are then published online, where users can quickly have access to neighboring State counts by querying the database base through a web browser based on year and/or week number; a simple mocked up site demonstrating this possibility can be seen in Figure 5.19.



localhost/logintaverna.php

Username : kdavis

Password : ••••••

localhost/toprocessauto.php

logged in as kdavis

Data Available from Taverna Workflow

(Data are available from Week 1 2003 through Week 20 2013)

Year: 2012 Week*: 02

localhost/processauto.php

Week	Year	City	State	Count
2	2012	Salt Lake City	Utah	5
2	2012	Salt Lake City	Utah	5
2	2012	Los Angeles	California	11
2	2012	Las Vegas	Nevada	4
2	2012	Los Angeles	California	11
2	2012	Las Vegas	Nevada	4

Figure 5.19: Mocked website

Discussion of Outcomes

An in-depth discussion of the results are presented in Chapter 6, however, to highlight significant findings, the following points are highlighted:

1. The quantitative analysis using more sophisticated natural language processing techniques to standardize and code death certificates provided the first stepping stones on the value of applying this technology to other public health data. However, some main problems when MetaMap performed below average need to be highlighted:
    - There are multiple concepts (noun phrases) in a single string
    - The phrase has a compound statement ("metastasis to brain and bone" or "gunshot wounds of the head and right arm")
    - The phrases begin with certain words (i.e., complications, etc...)
2. The results of the feasibility study clearly showed that grid technology can be used to help create a decentralized surveillance architecture. Although there are many grid middleware systems, for this project, we focused on caGrid. While the tools provided allowed us to develop the grid services, it is important to note that the ease of developing grid services using caGrid heavily depends on one's operating system. The grid services were developed on both Windows 7 operating system and Ubuntu, and we found that that it was easier to develop on Ubuntu, particularly analytical services. Under a Windows environment, the developer has to change the generated Java codes for when uploading, accessing, and transferring files. This is because the generated Java code is for an UNIX environment; therefore, the developer has to change all codes related to file

handling to include "\\" or "/."

Summary

This chapter provided the results obtained from the two phases of this research study. The first phase of the research involved quantitative analysis of comparing the current method of identifying pneumonia and influenza deaths to the proposed pipeline, composed of a detection rule and natural language processor, for the real-time encoding of death certificates using the identification of pneumonia and influenza cases. Recall, precision, positive predictive value, and F-measure were calculated and the results were presented showing that the DCP performed statistically better than the keyword-based system. In the second phase of this research project, we focused on demonstrating the use of grid technology to aid in surveillance practices. The results of this feasibility study were validated through a series of use case scenarios. The outcomes from all of these are used to demonstrate how current technology can be used to address gaps and areas of weakness in the surveillance architecture, more specifically mortality surveillance, and to make recommendations to decrease these deficiencies, all of which are dealt with in the next chapter.

CHAPTER 6

DISCUSSION

The final stage of this investigation involves offering general observations for each phase of the study, as well as their limitations. The study's relevance to biomedical informatics is also discussed.

General Observations Phase 1

To the best of my knowledge, this research project is the first to publish and disseminate findings on using a natural language processing tool and the UMLS to code and standardized death certificates to identify pneumonia- and influenza-related cases. The Death Certificates Pipeline (DC) developed here was statistically better than keyword searching and was computationally efficient. Moreover, through this study, we were able to show that the current method used by the Utah Department of Health underestimated pneumonia and influenza deaths in Utah. The simple keyword search method not only decreased recall and precision but also reduced the level of agreement. When reporting counts for surveillance purposes, one should be as accurate as possible; however, there is a tradeoff between recall and precision. A respectable recall value helps capture "true" pneumonia- or influenza-related deaths, while a respectable precision value is needed to avoid overestimating the number of weekly pneumonia-influenza deaths. For disease surveillance, increased precision enables public health officials to more accurately focus resources on the proper control and prevention measures. As a result, although both methods in this study had good precision values (DCP: 0.98 and keyword searching: 0.96), using the pipeline developed in this dissertation would be more advantageous. The sample size for this phase of the study was sufficient to show the difference between the two methods.

Although the MetaMap component was not evaluated, based on the DCP's output, one can hypothesize that it did an excellent job at extracting and processing cause of deaths text; this is consistent with the results of Reid et al. [106]. The performance of this (DCP) system is determined largely by the coverage of terms and sources in the UMLS. The comprehensiveness of the UMLS resulted in most of the concepts on the death certificates being present in the UMLS, which attributed to the good recall (0.998). The DCP's weakest strength was its precision (0.98). As mentioned in the Results chapter (Chapter 5), most of the concepts the system did identify (n=9) and for the keyword searching method (n=9) were due to data entry errors. Standardizing the data entry for pneumonia negation text such as "aspiration pneumonia" would overcome both methods' limitation of falsely identifying these death literals as pneumonia-related deaths.

In terms of timing, keyword searching was the faster method; however, the DCP was also in the sub 1/10 second range, thus implying that if the study's current hardware and configurations are used, one can process Utah's daily death (~40) in approximately 5.47 seconds and all deaths in the US (~ 6646) in approximately 909.17 seconds. This timing would be much faster than the minimum of two weeks to receive the coded data from the current CDC process. Moreover, these timings make it apparent that this system can be integrated in a real-time surveillance system without introducing any additional bottlenecks.

General Observations Phase 2

The main goal of this phase of the study is to investigate the use of grid technology to replace current processes in public health surveillance**.** Although the traction of discussing the role of grid technology as it applies to public health is slowly dwindling. This goal was achieved by describing grid technology methods and demonstrating the potential uses of the proposed method as exemplified through a series of use case scenarios and prototypes, thus demonstrating the value of this technology to the public health domain. Although the grid services are not ready to be implemented within the public health setting, it serves as a proof-of-concept implementation that can be perfected with additional time and software development investment. However, the results for this study provided five concrete deliverables:

1. The ability to create public health grids provides access to disparate data sources and allow for distributed analytics (such as natural language processing).
2. The use of the "mortality surveillance" workflow provides a framework for demonstrating how the public health grid can enhance current processes.
3. Articulating the steps to migrate data and analytical tools into a grid informs researchers and practitioners of these solutions and possible challenges.
4. The use of secure web services to enable access to previously isolated data sources for biosurveillance consumption is an emerging and appealing concept. This model aligns itself with ongoing, existing efforts within healthcare IT domain such as the Nationwide Health Information Network (NHIN). NHIN's

vision of providing healthcare information exchanges among hospitals, laboratories, and independent healthcare providers is based upon federated data access and the use of secure web services in accordance with Web Services Interoperability Organization (WS-I) standards.

5. Proof-of-concept studies demonstrated how public health organizations can combine multiple grid services to create a standardized workflow for biosurveillance and/or other public health-related purposes.

Although this phase of the study focused on the development of grid services related to mortality surveillance, this model can be applied to other data and analytical services, with the potential of providing valuable biosurveillance information. However, based on our experience, we believe that it is possible to create a platform where public health departments contribute different services to the grid. It is important to note that while modeling data services are more generalizable, when determining what services to deploy as a grid service, developers should first determine what public health problem this service will address; and secondly, they should create services that are both generalizable and specific to different public health functions.

In 2007, the then CDC National Center for Public Health Informatics started the initiative of creating a public health grid (PHGrid) [196] to make data readily available to various stakeholders. During this initiative, the CDC and its collaborators planned carrying out a series of proof-of-principle tests to help determine the benefits and advantages of using grid computing in the public health domain. Through the lessons learned from these experiments, the CDC was able to develop a pilot project that enabled secure and timely exchange of information across multiple areas [196]. Although the exchange of information is essential to enhancing current public health surveillance architectures, it is important to demonstrate the use of grid technology to not only share information but also to provide access to analytical services that can aid in developing a more complex grid application, concatenating these analytic services with appropriate data sources. This phase of this study produced a series of proof-of-concepts to address this gap by creating both data and analytical services and a workflow uses these services to model a current public health surveillance process, mortality surveillance. As a result, this study has the potential to pave the way for the future**.**

Assumptions

There are a few assumptions underlying my research proposal. This proposal is seeking to introduce new technologies into a public health environment. This intervention carries the assumptions that:

1. Most health departments do not have sufficient technical capacity to enhance their biosurveillance capabilities.
2. The use cases will be sufficient to demonstrate the potential benefits associated with leveraging technologies to enhance public health surveillance.
3. The death certificates utilized in Aims 1 and 3 are complete and accurate enough for public health case identification and identification of possible emerging diseases.

Limitations

Although the biomedical informatics has been around for many years, applying informatics to the public health field (public health informatics) is relatively new;

therefore, there is no defined methodology for performing the studies in this dissertation. As a result, the study implemented methodologies from different disciplines and general guidelines were followed to ensure statistical and methodological rigor. Despite these precautions, this study is still affected by several limitations; this section discusses these issues.

The death data used throughout this study were only from one institution. While efforts are being made to ensure consistency of death data across states, the way the data are captured may differ by states. As a result, the UML models created in this study, which are modeled after the Utah Health Department, cannot be generalizable to other states. Nevertheless, we have provided a starting point for creating data grid services for the public health data.

The pipeline created in the first phase of the study demonstrated that natural language processing tools can be used to identify pneumonia and influenza deaths. However, the performance of this pipeline is affected by the limitations of the used knowledge base (UMLS) and concept-recognition system (MetaMap). While a number of steps can be developed to improve the pipeline's shortcomings, the UMLS and MetaMap limitations are out of reach of the potential developers of a system. In addition, an evaluation of the natural language processing component of the death certificates pipeline was performed; as a result, further research is needed to determine the performance of a natural language processing tool specifically to the coding of death certificates. Further research also needs to be carried out to examine the use of the DCP on electronic death records across institutions and countries that may have different documentation procedures.

Although the grid services were validated through a series of proof-of-concept studies, the practical gains for the analytical services in terms of computational power was not done. In addition, a formal user acceptance testing was not performed; nevertheless, we believe that this work provides the foundation to aid researchers and developers in using grid technology to access distributed data and analytical services.

Overall Significance

Public health and medical information sharing have evolved greatly over the past decade. There has been a significant amount written in regards to the need for enhancing public health surveillance capabilities, specifically as a mechanism to increase situational awareness and common operating picture. In particular, the literature addresses the need for an integrated national biosurveillance enterprise and methods to identify emerging diseases. However, the literature is lacking prototypes or resources that cash-stricken public health departments can utilize as models to provide such an infrastructure. This research seeks to bridge gaps between information technology development and public health surveillance practice that prevent development of effective infrastructure for the sharing of information and techniques from being adequately utilized and evaluated to enhance public health surveillance. Each section of this research shows how to implement technologies that can significantly lower the barrier toward transforming public health surveillance systems, and provides a more specific framework for implementing nontraditional approaches to address the current gap in our nation's surveillance capabilities. Moreover, by providing a set of examples to demonstrate that these technologies benefit public health may provide insight that may lead to a better, more collaborative system of tools and datasets that will become the CDC's public health surveillance of the future.

Relevance to Biomedical Informatics

Public health informatics (PHI) is a subdomain of biomedical informatics, and like biomedical informatics, PHI is an interdisciplinary field that applies information science, computer science, statistics, and technology to overcome public health problems and improve public health processes [49]. Accordingly, the purpose of this study was to help solve ongoing public health problems**.** To address these problems, scientific tools were developed and techniques were leveraged; it is hoped that such knowledge will serve as building blocks for future research and ultimately have a positive impact on human health. Moreover, due to the interdisciplinary nature of the biomedical/public health informatics field, the necessary tools and knowledge needed to undertake this study was provided. This dissertation has also contributed to the field of biomedical and public health informatics:

- It demonstrated the effectiveness of natural language processing tools for processing unstructured public health data, thus giving insight on exploration of new methods to enhance how the healthcare community transforms, manipulates, and analyzes unstructured data sources.
- It enabled R and 'MetaMap' as analytical services in the caBIG environment. In the biomedical informatics community, MetaMap and R are commonly used and a de facto standard. However, to our knowledge, there has not been any attempt for grid enabling the MetaMap environment itself or R and using it as user interface for accessing the grid service. This dissertation provided insight into how a grid

service for these tools can be advantageous to the field of public health informatics.

- It provided a step-by-step process for creating caGrid data and analytical services, which can help developers who are having difficulties creating these services with the caGrid suite of tools. In addition, the use of the Taverna workflow engine showed how to integrate these services to simulate a current public health surveillance workflow; the methods used can be generalized to many other similar surveillance workflows in the public health domain.

CHAPTER 7

CONCLUSIONS

This project aimed to provide an overview of the current challenges faced by current public health surveillance systems. Key problem areas were identified and informatics techniques were leveraged to demonstrate that these methods can be used to overcome key challenges. Moreover, this works illustrates the following:

- This study shows that it is feasible to achieve high levels of accuracy when using NLP tools to identify cases of pneumonia and influenza cases from electronic death records while still providing a system that can be used for real-time coding of death certificates. Identification of concept identifiers related to the CDC's case definition of pneumonia and influenza was very important in producing a highly accurate rule for the identification of these cases.
- We demonstrated that it is possible to easily deploy grid applications for public health surveillance use. We concluded that the techniques used could be generalized to any application that has a command line interface. We believe that by providing a set of examples demonstrating the benefit of this technology to public health surveillance may lead to a better, more collaborative system for public health surveillance.

The difficult road of bringing technology used in academia to the public health domain, particularly to overcome key challenges faced by the public health surveillance community, still lies ahead. While this project is a first step to providing key insight and a roadmap on how new technologies in different fields can be used to enhance public health surveillance capabilities and infrastructure, further research is needed. Moreover, informatics tools and techniques such as those presented in this study promise to guide researchers and public health informatics professionals at current state and federal agencies in their efforts to improve public health practice through informatics.

APPENDIX A

IDENTIFICATION OF PNEUMONIA AND INFLUENZA DEATHS USING THE DEATH CERTIFICATES PIPELINE

Citation: Davis K, Staes C, Duncan J, Igo S, Facelli JC: **Identification of pneumonia and influenza deaths using the death certificate pipeline**. *BMC Medical Informatics and Decision Making* 2012, **12**(1):37

Davis *et al. BMC Medical Informatics and Decision Making* 2012, **12**:37
http://www.biomedcentral.com/1472-6947/12/37

BMC
Medical Informatics & Decision Making

RESEARCH ARTICLE **Open Access**

Identification of pneumonia and influenza deaths using the death certificate pipeline

Kailah Davis[1], Catherine Staes[1], Jeff Duncan[2], Sean Igo[1,3] and Julio C Facelli[1,3*]

Abstract

Background: Death records are a rich source of data, which can be used to assist with public surveillance and/or decision support. However, to use this type of data for such purposes it has to be transformed into a coded format to make it computable. Because the cause of death in the certificates is reported as free text, encoding the data is currently the single largest barrier of using death certificates for surveillance. Therefore, the purpose of this study was to demonstrate the feasibility of using a pipeline, composed of a detection rule and a natural language processor, for the real time encoding of death certificates using the identification of pneumonia and influenza cases as an example and demonstrating that its accuracy is comparable to existing methods.

Results: A Death Certificates Pipeline (DCP) was developed to automatically code death certificates and identify pneumonia and influenza cases. The pipeline used MetaMap to code death certificates from the Utah Department of Health for the year 2008. The output of MetaMap was then accessed by detection rules which flagged pneumonia and influenza cases based on the Centers of Disease and Control and Prevention (CDC) case definition. The output from the DCP was compared with the current method used by the CDC and with a keyword search. Recall, precision, positive predictive value and F-measure with respect to the CDC method were calculated for the two other methods considered here. The two different techniques compared here with the CDC method showed the following recall/ precision results: DCP: 0.998/0.98 and keyword searching: 0.96/0.96. The F-measure were 0.99 and 0.96 respectively (DCP and keyword searching). Both the keyword and the DCP can run in interactive form with modest computer resources, but DCP showed superior performance.

Conclusion: The pipeline proposed here for coding death certificates and the detection of cases is feasible and can be extended to other conditions. This method provides an alternative that allows for coding free-text death certificates in real time that may increase its utilization not only in the public health domain but also for biomedical researchers and developers.

Trial Registration: This study did not involved any clinical trials.

Keywords: Public health informatics, Natural language processing, Surveillance, Pneumonia and influenza

Background

The ongoing monitoring of mortality is crucial to detect and estimate the magnitude of deaths during epidemics, emergence of new diseases (for example, seasonal or pandemic influenza, AIDS, SARS), and the impact of extreme environmental conditions on a population such as heat waves or other relevant public health events or threats [1,2]. The surveillance of vital statistics is not a novel idea; mortality surveillance has played an integral part in public health since the London Bills of Mortality were devised in the seventeenth century [3]. The Bills served as an early warning tool against bubonic plague by monitoring deaths from the 1635 to the 1830s. Today, mortality surveillance continues to be a critical activity for public health agencies throughout the world [4-7].

Pneumonia and influenza are serious public health threats and are a cause of substantial morbidity and mortality worldwide; for instance, the World Health Organization (WHO) estimates seasonal influenza causes between

* Correspondence: Julio.facelli@utah.edu
[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA
[3]Center for High Performance Computing, University of Utah, Salt Lake City, Utah, USA
Full list of author information is available at the end of the article

250,000 to 500,000 deaths worldwide each year [8] while pneumonia kills more than 4 million people worldwide every year [9]. Worldwide, the morbidity and mortality of influenza and pneumonia have a considerable economic impact in the form of hospital and other health care costs. Each year in the United States approximately 3 million persons acquire pneumonia and, depending on the severity of the influenza season, 15 to 61 million people in the US contract influenza [9]. These numbers contribute to approximately 1.3 million hospitalizations, of which 1.1 million are pneumonia cases [10] and the remainder for influenza [11]. Moreover, pneumonia cases and influenza together cost the American economy 40.2 billion dollars in 2005 [12]. In The Netherlands it has been estimated that influenza accounts for 3713 and 744 days of hospitalization per 100,000 high-risk and low-risk elderly, respectively [13]. Due to the public health burden and the unpredictability of an influenza season, strong pneumonia and influenza surveillance systems are a priority for health authorities.

Mortality monitoring is an important tool for the surveillance of pneumonia and influenza which can aid in the rapid detection and estimates of excess deaths and inform and evaluate the effect of vaccination and control programs. Traditionally, influenza mortality surveillance often uses the category of "pneumonia and influenza" (P-I) on death certificates as an indicator of the severity of an influenza season or to identify trends within a season; however, only a small proportion of these deaths are influenza related. It has been reported that only 8.5–9.8% of all pneumonia and influenza deaths are influenza related [14,15]. The non-influenza-related pneumonia deaths tend to be stable from year to year and fluctuations in this category are largely driven by the prevalence and severity of seasonal influenza. As a result, the P-I category is an important sentinel indicator.

In the US, death certificates are the primary data source for mortality surveillance whose findings are widely used to exemplify epidemics and measure the severity of influenza seasons [16]. Currently, there are three systems to monitor influenza-related mortality; one system in particular, the 122 Cities Mortality Reporting System, provides a rapid assessment of pneumonia and influenza mortality [6]. Each week, this system summarizes the total number of death certificates filed in 122 US cities, as well as the number of deaths due to pneumonia and influenza. However, even these data can be delayed by approximately 2–3 weeks from the times of death. This delay can be attributed to one of the following reasons: 1) timeliness of death registration and 2) reviewing of the death certificates to identify pneumonia and influenza deaths [6,16,17]. The registration and reviewing of death certificates varies by states and, as a result, there is variability in length of time to report a death to CDC. For instance, states with paper-based death registration system typically perform manual reviews of the death certificates which can take up to 3 weeks; however states with electronic death registration systems (EDRS) may perform automatic reviews which can decrease this time significantly.

The current 122 Cities Mortality Reporting System surveillance system also lacks flexibility for expanding the number of conditions and/or the geographic distribution. Moreover, the unavailability of coded death records due to the complexity of the National Center of Health Statistics (NCHS) coding process results in multiple strategies to identify common outbreaks such as pneumonia and influenza deaths, which greatly vary by jurisdiction. To bypass the lengthy NCHS process, a variety of approaches have been attempted that are close to 'real- time' but less than optimal. For instance, in Utah keyword searching is used to identify pneumonia and influenza deaths; although this method is fast and easy to implement, it can easily result in the over or under estimation of cases. This can occur by missing cases due to misspelled terms, synonyms, variations, or the selection of strings containing the search term.

Other research groups [18,19] have demonstrated the feasibility of using mortality data for real time surveillance but all used "free text" search for the string "pneumonia", "flu" or "influenza." As noted earlier, although this method can provide the semi quantitative measurements for disease surveillance purposes, keyword searches can also result in an array of problems that result from complexities of human language such as causal relationships and synonyms [20]. Therefore, the lack of coded death data that may not be available for months [21] seriously limits the use of death records in automated systems. At this time, there is little published on the automatic assignment of codes to death certificates for automatic case detection.

Coding death certificates

Currently the coding of death certificates is a complex process which involves many entities. In the US, where we are focusing this study, the codes on death certificates that are generated by the National Center for Health Statistics (NCHS) **depend on information reported on the death certificate** by the medical examiner, coroner, or another certifier, and there is substantial variation in how certifiers interpret and adhere to cause-of-death definitions [22]. The cause of death literals are coded into International Classification of Diseases Tenth Revision (ICD-10) [23] and the underlying and multiple-cause-of-death codes are selected based on the World Health Organization coding rules. These coding rules have been automated by CDC with the development the Mortality Medical Data System (MMDS) which consists of four programs: Super Mortality Medical Indexing Classification and Retrieval (SuperMICAR) Data Entry; Mortality Medical Indexing Classification and Retrieval (MICAR); Automated Classification of Medical Entities (ACME) and TRANSAX (Translation Axes). SuperMICAR was designed to facilitate the entry of literal text of causes of death in death certificates and convert them into standardized expressions acceptable by MICAR. It contains a dictionary which assigns an entity reference number (ERN) to statements on the death certificate. These ERNs are fed into MICAR200 which transforms the ERNS into ICD-10 codes by using specific mortality coding rules; the rules require look-up files and a dictionary. ACME and TRANSAX then selects the underlying and multiple causes of death respectively. ICD-10 codes from MICAR200 are fed into ACME which assigns the underlying cause of death using decision tables. The decision table contains all possible pairings of diseases for which the first disease can cause the second. In the latest version of the system, ACME is comprised of eight decision tables including three tables of valid and invalid codes, causal relationships (General Principle and Rule 1), and direct sequel (Rule 3), and three other tables needed by modification rules. Figure A.1 provides the workflow for the MMDS system.

Of the 2.3 million deaths that occur each year 80–85 percent are automatically coded through SuperMICAR, and the remaining records are then manually coded by nosologists, a medical classification specialist [24]; this is a tedious and lengthy process lasting up to 3 months. Although the automation process has decreased the time required for coding death data to 1–2 weeks, the national vital statistics data is not available for at least two years. Therefore, local health department still manually code records or perform basic process techniques to quickly characterize disease patterns [25].

Records that were processed through Super-MICAR or were manually coded are then processed through the remaining components (MICAR200, ACME and TRANSAX) of MMDS. In 1999, MICAR200 had a throughput rate of 95–97%, while ACME rate was 98 percent. Moreover, based on a reliability study, ACME error rate for selecting the underlying cause is at one-half percent, while TRANSAX, the multiple cause codes had a one-half percent error rate [26]. Due to the high processing rates and low error rates, MMDS is considered by practitioners as the gold standard for the processing and coding of death certificates in the US and other countries (such as Canada, the United Kingdom (UK) and Australia). Therefore, we used the codes produced by this system as the "gold standard" when comparing with the methods developed here.

Electronic death registration system

In 1997, the US Steering Committee to Reengineer the Death Registration Process (a task force representing federal agencies, the National Center for Health Statistics and the Social Security Administration, and professional organizations representing funeral directors, physicians, medical examiners, coroners, hospitals, medical records professionals, and vital records and statistics officials (NAPHSIS) published the report "*Toward an Electronic Death Registration System in the United States: Report of the Steering Committee to Reengineer the Death Registration Process.*" This report explained the feasibility of developing electronic death registration in the United States [27] and argued that these electronic death records have the potential to be an effective source of information for nation-wide tracking and detecting of disease outbreaks. However, little actions have been taken to implement such recommendations in a comprehensive manner. As of July 2011, electronic death registration systems were operating in 36 states, the District of Colombia, and in development or planning stage in a dozen others [28].

NLP potential

Information representing the 'cause of death' field on the death certificates is free text. One major goal of natural language processing (NLP) is to extract and encode data from free- texts. There have been many research groups developing NLP systems to aid in clinical research, decision support, quality assurance, the automation of encoding free text data and disease surveillance [29-31]. Although, there have been a few NLP applications to the public health domain [32,33], little is known about its capability to automatically code death certificates for outbreak and disease surveillance. Recently, Medical Match Master (MMM) [25], developed by Riedl et al at the University of California Davis, was used to match unstructured cause of death phrases to concepts and semantic types within the Unified Medical Language System (UMLS). The system annotates each death phrase input with two types of information, the Concept Unique Identifier, CUI, and a semantic type both assigned by the UMLS. MMM was able to identify an exact concept identifier (CUI) from the UMLS for over 50% of 'cause of death' phrases. Although, the focus of this study was to use NLP techniques to process death certificates, the description of this system reported in the literature did not show how well coded data from an NLP tool along with predefined rules can detect countable cases for a specific disease or condition.

The purpose of our project is to create a pipeline which automatically encodes death certificates using a NLP tool and identify deaths related to pneumonia and influenza which provides daily and/or weekly counts. We compared the new technique developed here with

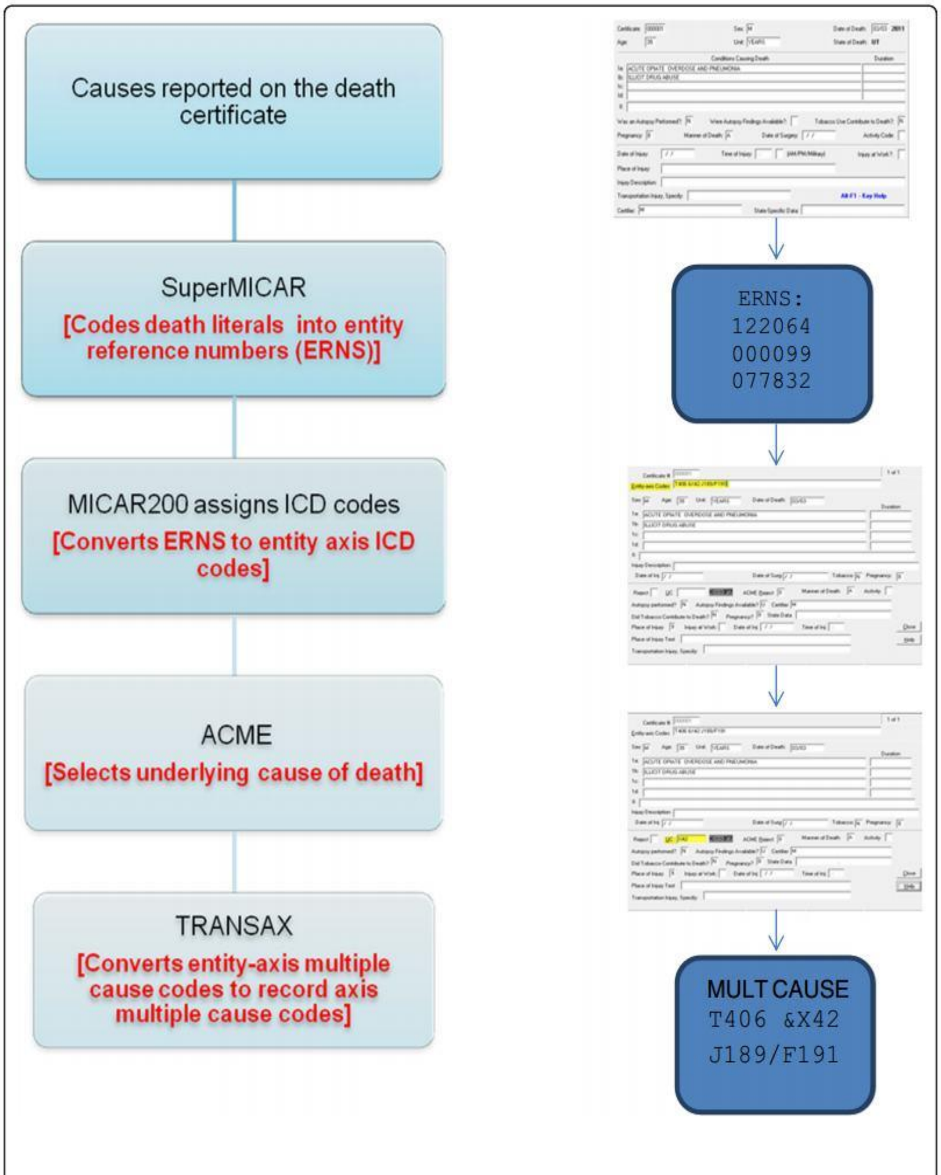

Causes reported on the death certificate
SuperMICAR
[Codes death literals into entity reference numbers (ERNS)]
MICAR200 assigns ICD codes
[Converts ERNS to entity axis ICD codes]
ACME
[Selects underlying cause of death]
TRANSAX
[Converts entity-axis multiple cause codes to record axis multiple cause codes]
ERNS:
122064
000099
077832
MULT CAUSE
T406 &X42
J189/F191

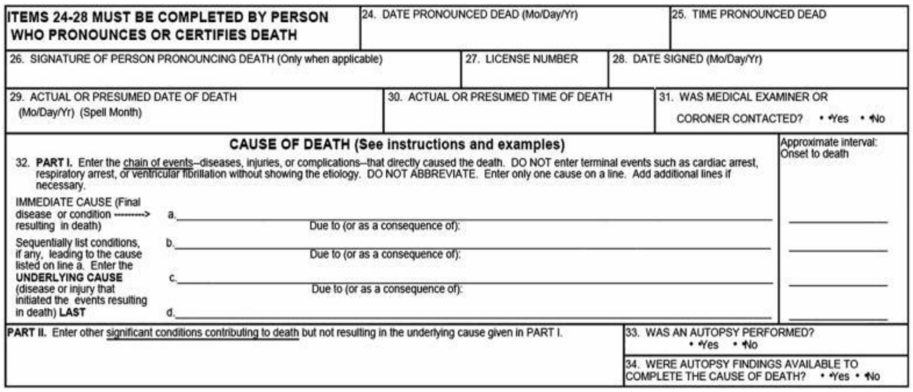
Figure A.1 The Mortality Medical Data System (MMDS) Workflow.

keyword searching and MMDS as exemplars of the easiest possible approach and the current "gold standard", respectively. The comparison of the techniques was done by calculating recall, precision, F- measure, positive predictive value and agreement (Cohen's Kappa).

Methods

Sample

We obtained 14,440 de-identified electronic death records all with multiple-cause-of-death from the Utah Department of Health (UDOH) for the period 1



ITEMS 24-28 MUST BE COMPLETED BY PERSON WHO PRONOUNCES OR CERTIFIES DEATH

24. DATE PRONOUNCED DEAD (Mo/Day/Yr)

25. TIME PRONOUNCED DEAD

26. SIGNATURE OF PERSON PRONOUNCING DEATH (Only when applicable)

27. LICENSE NUMBER

28. DATE SIGNED (Mo/Day/Yr)

29. ACTUAL OR PRESUMED DATE OF DEATH
(Mo/Day/Yr) (Spell Month)

30. ACTUAL OR PRESUMED TIME OF DEATH

31. WAS MEDICAL EXAMINER OR CORONER CONTACTED? • Yes • No

CAUSE OF DEATH (See instructions and examples)

Approximate interval: Onset to death

32. **PART I.** Enter the chain of events--diseases, injuries, or complications--that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.

IMMEDIATE CAUSE (Final disease or condition ---------> resulting in death) a.______________ ______
Due to (or as a consequence of):

Sequentially list conditions, if any, leading to the cause listed on line a. Enter the **UNDERLYING CAUSE** (disease or injury that initiated the events resulting in death) **LAST**

b.______________ ______
Due to (or as a consequence of):

c.______________ ______
Due to (or as a consequence of):

d.______________ ______

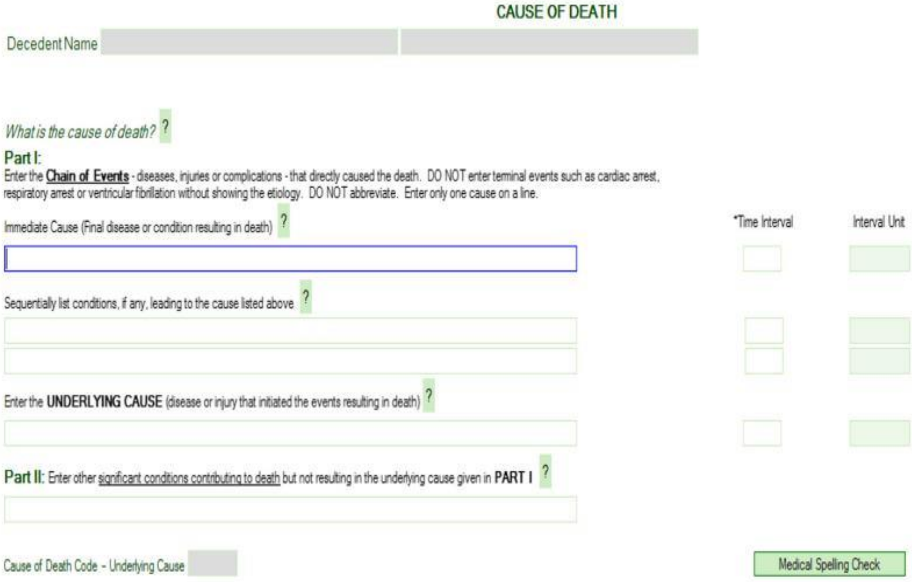
PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I.

33. WAS AN AUTOPSY PERFORMED?
• Yes • No

34. WERE AUTOPSY FINDINGS AVAILABLE TO COMPLETE THE CAUSE OF DEATH? • Yes • No

Figure A.2 Portion of the US standard certificate of death in which cause-of-death data are entered.

January 2008 to 31 December 2008. The records included a section describing the disease or condition directly leading to death, and any antecedent causes, co-morbid conditions and other significant contributing conditions. An example of a paper and electronic death certificate are shown in Figures A.2 and A.3 respectively. All death certificates used in this study have been processed using the Mortality Medical Data System (MMDS) and the record axis codes were received from UDOH.



CAUSE OF DEATH

Decedent Name

What is the cause of death? ?

Part I:
Enter the **Chain of Events** - diseases, injuries or complications - that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest or ventricular fibrillation without showing the etiology. DO NOT abbreviate. Enter only one cause on a line.

Immediate Cause (Final disease or condition resulting in death) ?

*Time Interval

Interval Unit

Sequentially list conditions, if any, leading to the cause listed above ?

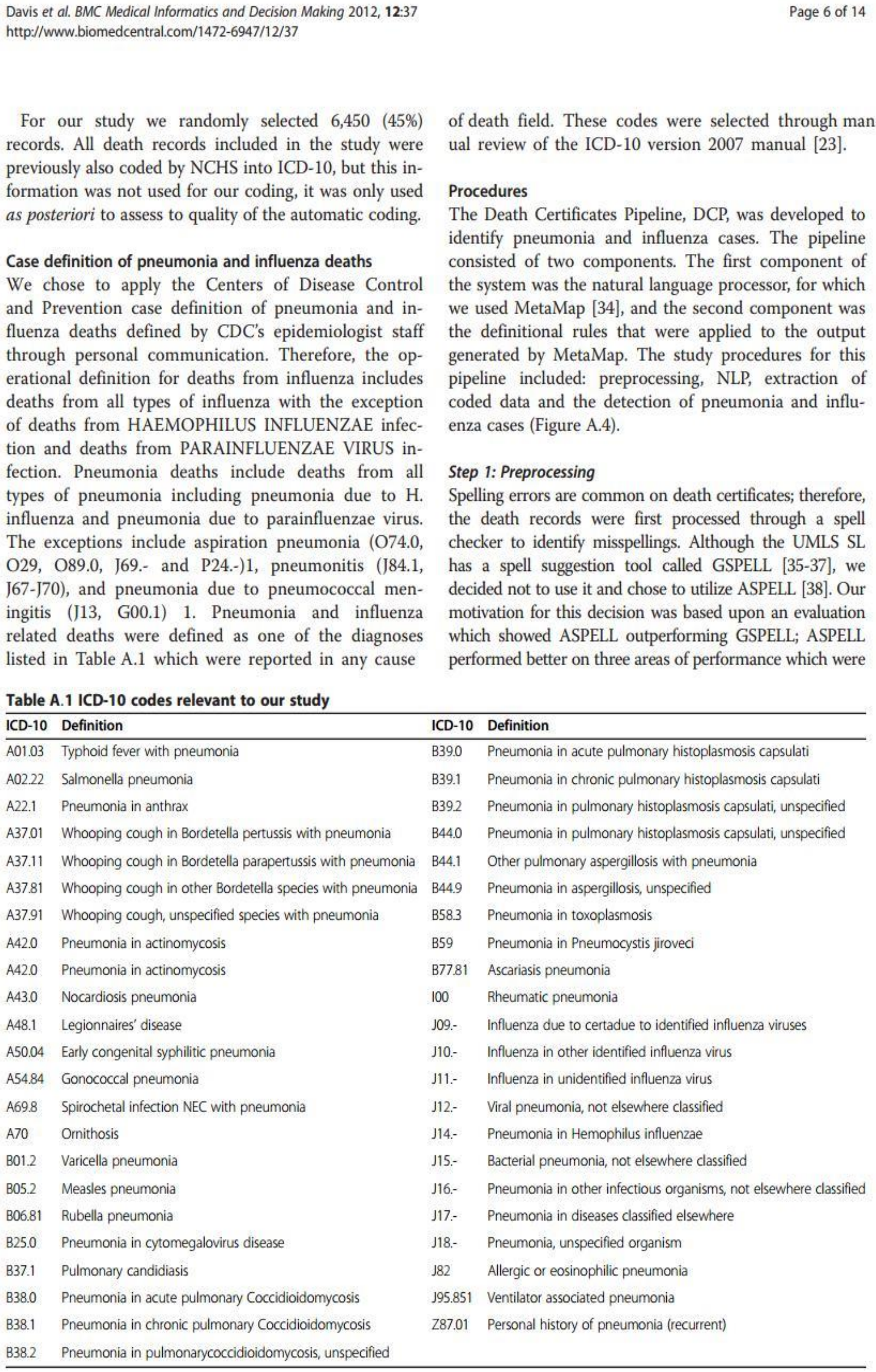
Enter the **UNDERLYING CAUSE** (disease or injury that initiated the events resulting in death) ?

Part II: Enter other significant conditions contributing to death but not resulting in the underlying cause given in **PART I** ?

Cause of Death Code - Underlying Cause

Medical Spelling Check

Figure A.3 Utah department of health electronic death certificate.



For our study we randomly selected 6,450 (45%) records. All death records included in the study were previously also coded by NCHS into ICD-10, but this information was not used for our coding, it was only used *as posteriori* to assess to quality of the automatic coding.

Case definition of pneumonia and influenza deaths

We chose to apply the Centers of Disease Control and Prevention case definition of pneumonia and influenza deaths defined by CDC's epidemiologist staff through personal communication. Therefore, the operational definition for deaths from influenza includes deaths from all types of influenza with the exception of deaths from HAEMOPHILUS INFLUENZAE infection and deaths from PARAINFLUENZAE VIRUS infection. Pneumonia deaths include deaths from all types of pneumonia including pneumonia due to H. influenza and pneumonia due to parainfluenzae virus. The exceptions include aspiration pneumonia (O74.0, O29, O89.0, J69.- and P24.-)1, pneumonitis (J84.1, J67-J70), and pneumonia due to pneumococcal meningitis (J13, G00.1) 1. Pneumonia and influenza related deaths were defined as one of the diagnoses listed in Table A.1 which were reported in any cause of death field. These codes were selected through manual review of the ICD-10 version 2007 manual [23].

Procedures

The Death Certificates Pipeline, DCP, was developed to identify pneumonia and influenza cases. The pipeline consisted of two components. The first component of the system was the natural language processor, for which we used MetaMap [34], and the second component was the definitional rules that were applied to the output generated by MetaMap. The study procedures for this pipeline included: preprocessing, NLP, extraction of coded data and the detection of pneumonia and influenza cases (Figure A.4).

Step 1: Preprocessing

Spelling errors are common on death certificates; therefore, the death records were first processed through a spell checker to identify misspellings. Although the UMLS SL has a spell suggestion tool called GSPELL [35-37], we decided not to use it and chose to utilize ASPELL [38]. Our motivation for this decision was based upon an evaluation which showed ASPELL outperforming GSPELL; ASPELL performed better on three areas of performance which were

Table A.1 ICD-10 codes relevant to our study

ICD-10	Definition	ICD-10	Definition
A01.03	Typhoid fever with pneumonia	B39.0	Pneumonia in acute pulmonary histoplasmosis capsulati
A02.22	Salmonella pneumonia	B39.1	Pneumonia in chronic pulmonary histoplasmosis capsulati
A22.1	Pneumonia in anthrax	B39.2	Pneumonia in pulmonary histoplasmosis capsulati, unspecified
A37.01	Whooping cough in Bordetella pertussis with pneumonia	B44.0	Pneumonia in pulmonary histoplasmosis capsulati, unspecified
A37.11	Whooping cough in Bordetella parapertussis with pneumonia	B44.1	Other pulmonary aspergillosis with pneumonia
A37.81	Whooping cough in other Bordetella species with pneumonia	B44.9	Pneumonia in aspergillosis, unspecified
A37.91	Whooping cough, unspecified species with pneumonia	B58.3	Pneumonia in toxoplasmosis
A42.0	Pneumonia in actinomycosis	B59	Pneumonia in Pneumocystis jiroveci
A42.0	Pneumonia in actinomycosis	B77.81	Ascariasis pneumonia
A43.0	Nocardiosis pneumonia	I00	Rheumatic pneumonia
A48.1	Legionnaires' disease	J09.-	Influenza due to certadue to identified influenza viruses
A50.04	Early congenital syphilitic pneumonia	J10.-	Influenza in other identified influenza virus
A54.84	Gonococcal pneumonia	J11.-	Influenza in unidentified influenza virus
A69.8	Spirochetal infection NEC with pneumonia	J12.-	Viral pneumonia, not elsewhere classified
A70	Ornithosis	J14.-	Pneumonia in Hemophilus influenzae
B01.2	Varicella pneumonia	J15.-	Bacterial pneumonia, not elsewhere classified
B05.2	Measles pneumonia	J16.-	Pneumonia in other infectious organisms, not elsewhere classified
B06.81	Rubella pneumonia	J17.-	Pneumonia in diseases classified elsewhere
B25.0	Pneumonia in cytomegalovirus disease	J18.-	Pneumonia, unspecified organism
B37.1	Pulmonary candidiasis	J82	Allergic or eosinophilic pneumonia
B38.0	Pneumonia in acute pulmonary Coccidioidomycosis	J95.851	Ventilator associated pneumonia
B38.1	Pneumonia in chronic pulmonary Coccidioidomycosis	Z87.01	Personal history of pneumonia (recurrent)
B38.2	Pneumonia in pulmonarycoccidioidomycosis, unspecified		

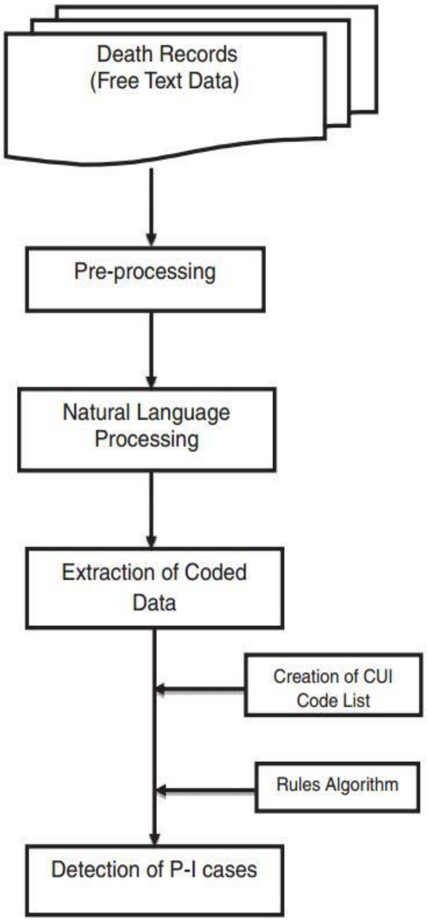


Figure A.4 Flow diagram of the death certificates pipeline.

evaluated: (1) whether the correct word was ranked number one; (2) whether the correct word was ranked in the top ten; and (3) whether the correct word was found at all [35]. PERL (www.perl.org), a high-level computer programming language that aids in the manipulation and processing of large volume of text data was then used to prepare the cause of death free text for NLP. The preprocessing also involved the removal of non-ASCII characters; this was a required technical step for MetaMap processing.

Step 2: Natural language processing

MetaMap was used to convert the electronic death records to coded descriptions appropriate for the rule based system. MetaMap [34], developed by the National Library of Medicine (NLM), is useful in identifying biomedical concepts from free-form textual input and maps them into concepts from the Unified Medical Language System (UMLS) Metathesaurus [34,39]. MetaMap works by breaking the inputted text into words or phrases, map them to standard terms, and then match the terms to concepts in the Unified Medical Language System (UMLS) [40]. For each matched phrase, MetaMap classifies it into a semantic type then returns the concept unique identifier (CUI) and the mapping options which are ranked according to the strength of the mapping. Table A.2 shows an example of sample death literal and its associated XML

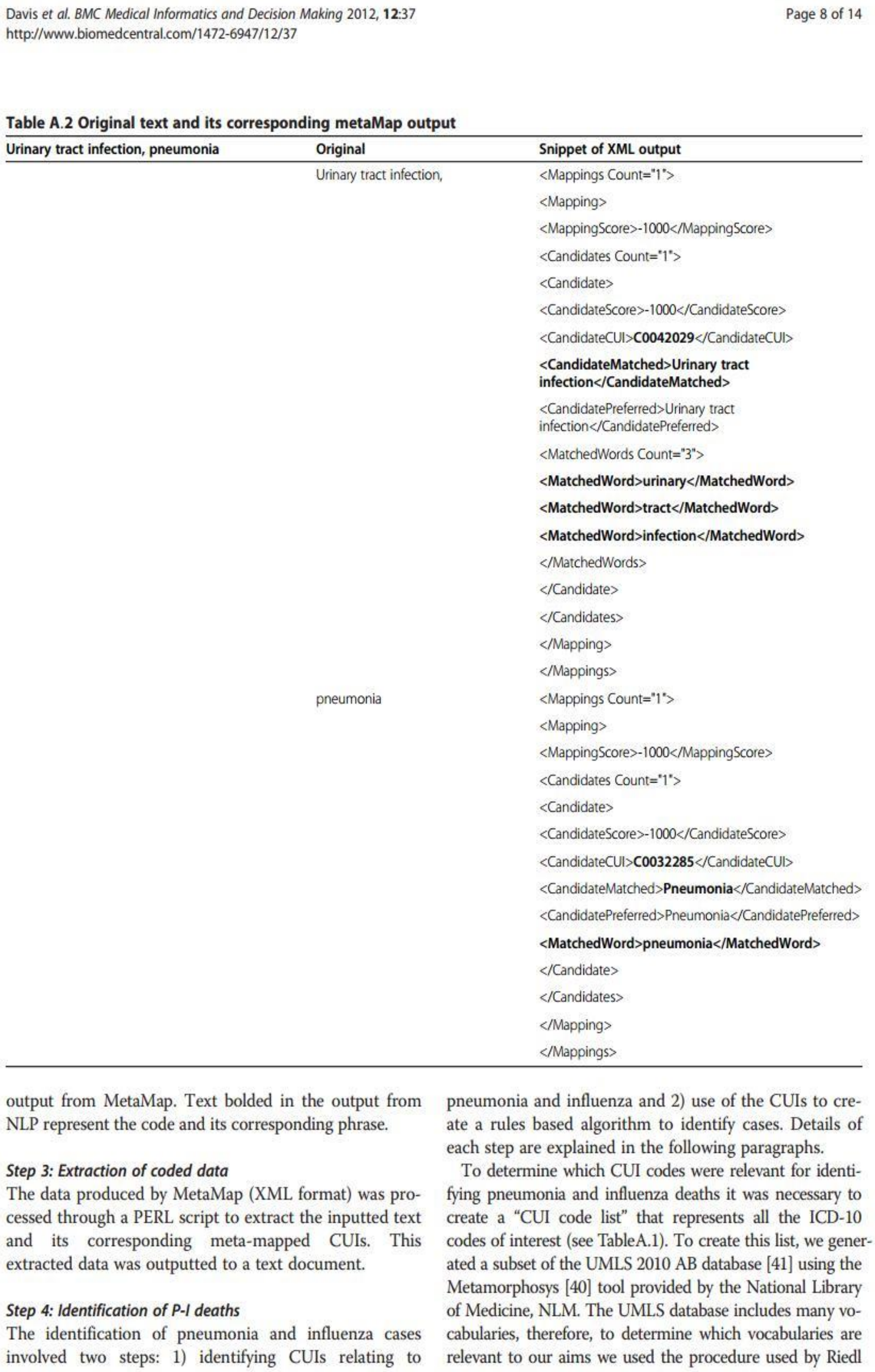


Table A.2 Original text and its corresponding metaMap output

Urinary tract infection, pneumonia	Original	Snippet of XML output
	Urinary tract infection,	<Mappings Count="1">
		<Mapping>
		<MappingScore>-1000</MappingScore>
		<Candidates Count="1">
		<Candidate>
		<CandidateScore>-1000</CandidateScore>
		<CandidateCUI>**C0042029**</CandidateCUI>
		<CandidateMatched>Urinary tract infection</CandidateMatched>
		<CandidatePreferred>Urinary tract infection</CandidatePreferred>
		<MatchedWords Count="3">
		<MatchedWord>urinary</MatchedWord>
		<MatchedWord>tract</MatchedWord>
		<MatchedWord>infection</MatchedWord>
		</MatchedWords>
		</Candidate>
		</Candidates>
		</Mapping>
		</Mappings>
	pneumonia	<Mappings Count="1">
		<Mapping>
		<MappingScore>-1000</MappingScore>
		<Candidates Count="1">
		<Candidate>
		<CandidateScore>-1000</CandidateScore>
		<CandidateCUI>**C0032285**</CandidateCUI>
		<CandidateMatched>**Pneumonia**</CandidateMatched>
		<CandidatePreferred>Pneumonia</CandidatePreferred>
		<MatchedWord>pneumonia</MatchedWord>
		</Candidate>
		</Candidates>
		</Mapping>
		</Mappings>

output from MetaMap. Text bolded in the output from NLP represent the code and its corresponding phrase.

Step 3: Extraction of coded data

The data produced by MetaMap (XML format) was processed through a PERL script to extract the inputted text and its corresponding meta-mapped CUIs. This extracted data was outputted to a text document.

Step 4: Identification of P-I deaths

The identification of pneumonia and influenza cases involved two steps: 1) identifying CUIs relating to pneumonia and influenza and 2) use of the CUIs to create a rules based algorithm to identify cases. Details of each step are explained in the following paragraphs.

To determine which CUI codes were relevant for identifying pneumonia and influenza deaths it was necessary to create a "CUI code list" that represents all the ICD-10 codes of interest (see TableA.1). To create this list, we generated a subset of the UMLS 2010 AB database [41] using the Metamorphosys [40] tool provided by the National Library of Medicine, NLM. The UMLS database includes many vocabularies, therefore, to determine which vocabularies are relevant to our aims we used the procedure used by Riedl *et al.* [25] which included every level 0 source plus SNOMED, a level 9 source. Sources such as National Center for Biotechnology Information (NCBI) taxonomy, which included ICD-10 codes, were also included in our configuration of the subset. For the purpose of our study, we focused on two tables within the UMLS schema. The first table MRCONSO contains a unique row for each lexical variant of a given concept. The second table MRREL contains information about the relationship among concepts. Tables A.3 and A.4 shows sample rows and columns in the MRCONSO and MRREL tables associated to CUIs rerelated to "pneumonia" [C0032285], "influenza" [C0021400] "pneumonia and influenza" [C0155870]. Our configuration produced 6,862,110 rows in MCRNOSO and 23, 467,822 rows in MRREL.

Three queries were performed on the subset described above to map pneumonia and influenza ICD-10 codes to CUIs and identify related pneumonia and influenza concepts. Each query was then placed in a separate database, all duplicates were removed and a sub-query was run to ensure that only the ICD-10 codes in TableA.1 were included in this list. This produced 241 distinct concept identifiers (CUIs) relating to pneumonia or influenza. These codes were used to develop the rules to identify the cases of interest.

The coded data produced by MetaMap was accessed by rules, aimed at identifying the presence of pneumonia and influenza based on the coded data. The rules for identifying these deaths used the CUI code list described above. The rule looks at each cause of death field (Underlying Cause, Additional Causes, etc.) to flag records with relevant codes. These rules used boolean operators (And, Or, Not) and if-then statements to create a chain of rules (Figure A.5).

Comparison methodologies

The list of cases identified by our automated detection system was compared with those identified by two other methods: a) keyword searching and b) the reference standard: the ICD-10 codes given by the CDC MMDS method. For key-word searching we followed the process utilized by the Utah Department of Health where all the cause of death fields were scanned for the text strings 'PNEUMONIA' OR 'INFLUENZA'. The words 'ASPIRATION PNEUMONIA', 'PNEUMONITIS', 'PNEUMOCOCCAL MENINGITIS', 'HAEMOPHILUS INFLUENZAE' and 'PARAINFLUENZAE VIRUS' were excluded.

To evaluate the performance of both techniques against the reference standard, we needed to specify what constituted a match. Each death record is associated to a unique number; therefore, we considered a match if the unique identifier was identified by the comparator and also found by the reference standard.

Statistical analysis

Three standard measures were used to evaluate the performance of one method in relation to the reference standard used in this study: precision (equivalent to positive predictive value; recall (equivalent to sensitivity or true positive rate), and F-measure. Kappa statistics were used to assess agreement and McNemar's test was used to analyze the significance between the two methods. All calculations were performed in R [42].

To calculate these values, pneumonia and influenza related deaths were examined by comparing the reference standard output vs. the two comparators: DCP and keyword search. For both comparators, the deaths were counted and categorized as TRUE POSITIVES (cases found by the comparator—pneumonia deaths being correctly classified); FALSE POSITIVES (incorrect cases found by the comparator—the number of pneumonia and influenza deaths incorrectly identified by the comparator); FALSE NEGATIVES (correct cases not found by the comparator—the number of pneumonia deaths not identified by the comparator). Precision, recall and F-score were calculated as follows:

Precision = True Positives/(True Positives + False Positives) (1)

Recall = True Positives/(True Positives + False Negatives) (2)

F-measure = 2 *(P R/ P + R) (3)

McNemar's test was also calculated to evaluate the significance of the difference between the two

Table A.3 Sample rows and columns from the MRCONSO table

CUI	LUI	SUI	AUI	SAUI	SCUI	SAB	CODE	STR
C0021400	L0021400	S0667823	A17788091	NULL	NULL	ICD10CM	J10.1	Influenza NOS
C0021400	L0016270	S0003527	A2875695	11205017	6142004	SNOMEDCT	6142004	Flu
C0021400	L0018238	S0046068	A2882094	11206016	6142004	SNOMEDCT	6142004	Grippe
C0032285	L3025870	S3482854	A15102770	2.76E + 09	60363000	SNOMEDCT	60363000	Pneumonia (disorder)
C0032285	L0880404	S0991677	A1049957	NULL	NULL	ICD10	J18.9	Pneumonia, unspecified
C0155870	L0182738	S1458440	A1411637	NULL	NULL	ICD10	J10-J18.9	Influenza and pneumonia
C0155870	L0182738	S0247321	A16973811	NULL	NULL	ICD9CM	487	Influenza with pneumonia

Table A.4 Sample rows and columns from the MRREL Table

CUI1	AUI1	STYPE1	REL	CUI2	AUI2	STYPE2	RELA	SAB	SL
C0021400	A0481781	AUI	RB	C0029342	A0318194	AUI	NULL	CSP	CSP
C0021400	A0481781	AUI	RN	C0276357	A1196494	AUI	NULL	CSP	CSP
C0021400	A0412457	AUI	RQ	C0021400	A0247343	AUI	mapped_from	CST	CST
C0032285	A0102675	SDUI	SIB	C0273115	A15577420	SDUI	NULL	MSH	MSH
C0032285	A18169362	CODE	RO	C0485207	A18252567	CODE	has_fragments for_synonyms	LNC	LNC
C0021400	A4386826	CODE	RB	C0348675	A0723374	CODE	mapped_from	ICPC2ICD10ENG	ICPC2ICD10ENG
C0032285	A1049957	AUI	PAR	C0339951	A0242105	AUI	NULL	ICD10	ICD10
C0155870	A1411637	AUI	CHD	C0339951	A0242105	AUI	NULL	ICD10	ICD10

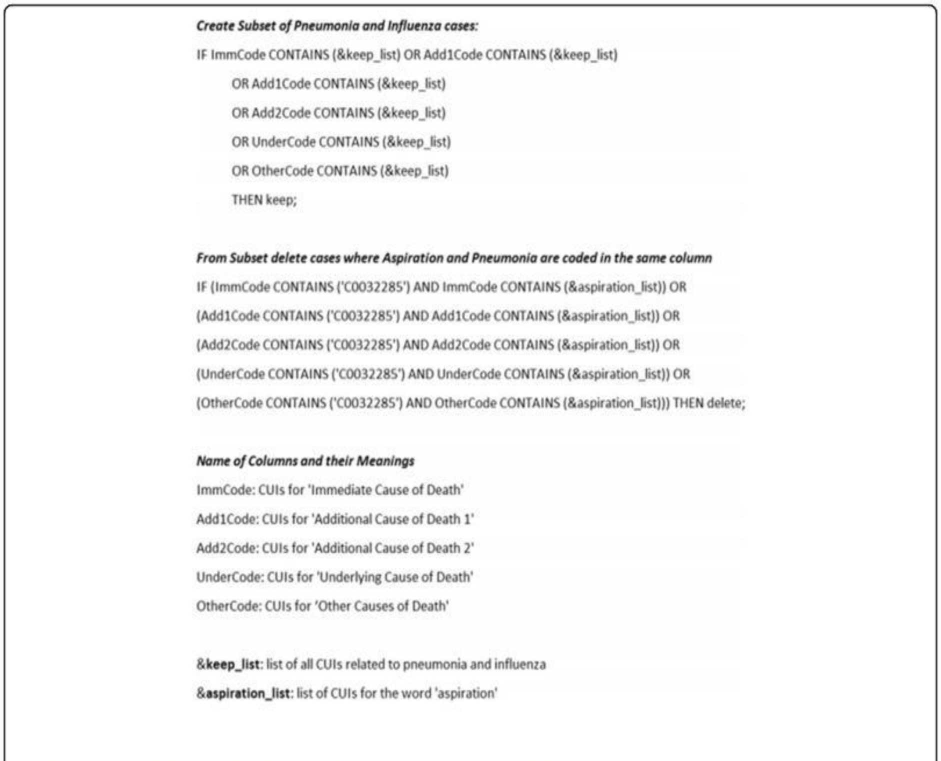
comparators. To calculate this value a confusion matrix was created where A is the number of times both methods have correct predictions; B is the number of times method 1 has a correct prediction and method 2 has a wrong prediction; C is the number of times method 2 has a correct prediction and method 1 has a wrong prediction; D is the number of times both methods have incorrect predictions.

Ethics approval was not required for this study. Identifying variables that could be used for re-identifying individuals were excluded from the study data.

Results

Processing time of the data

The records were processed and analyzed on a server with two Opteron Dual-Core 2.8 GHz processors and



```
Create Subset of Pneumonia and Influenza cases:
IF ImmCode CONTAINS (&keep_list) OR Add1Code CONTAINS (&keep_list)
     OR Add1Code CONTAINS (&keep_list)
     OR Add2Code CONTAINS (&keep_list)
     OR UnderCode CONTAINS (&keep_list)
     OR OtherCode CONTAINS (&keep_list)
     THEN keep;

From Subset delete cases where Aspiration and Pneumonia are coded in the same column
IF (ImmCode CONTAINS ('C0032285') AND ImmCode CONTAINS (&aspiration_list)) OR
(Add1Code CONTAINS ('C0032285') AND Add1Code CONTAINS (&aspiration_list)) OR
(Add2Code CONTAINS ('C0032285') AND Add2Code CONTAINS (&aspiration_list)) OR
(UnderCode CONTAINS ('C0032285') AND UnderCode CONTAINS (&aspiration_list)) OR
(OtherCode CONTAINS ('C0032285') AND OtherCode CONTAINS (&aspiration_list))) THEN delete;

Name of Columns and their Meanings
ImmCode: CUIs for 'Immediate Cause of Death'
Add1Code: CUIs for 'Additional Cause of Death 1'
Add2Code: CUIs for 'Additional Cause of Death 2'
UnderCode: CUIs for 'Underlying Cause of Death'
OtherCode: CUIs for 'Other Causes of Death'

&keep_list: list of all CUIs related to pneumonia and influenza
&aspiration_list: list of CUIs for the word 'aspiration'
```

Figure A.5 **Rules applied to MetaMap's output to extract pneumonia and influenza cases.**

16 GB RAM at the Center of High Performance Computing at the University of Utah. Using keyword searching the CPU processing time to identify pneumonia and influenza cases was 0.21 seconds and the wall time was 0.37 seconds. For the DCP, the total CPU processing time was 881.83 seconds. The NLP portion of the pipeline attributed to 99.4 percent of the processing time (NLP-877 seconds). While the DCP execution time is much longer, still it is well within the "in real time" realm. For instance, it would take 6,364.3 seconds CPU time seconds for DCP to code and flag all the weekly death records of the US (≈ 46,523).

Statistical analysis

Recall and precision were calculated at a 0.95 confidence intervals; the F-measure was also calculated. The performance of each method is described below.

Keyword searching

Of the 6,450 records analyzed keyword search identified 473 records as pneumonia and influenza deaths, 21 being identified as false positives. Precision for keyword searching was calculated at 96%. Of the 21 false positives, 6 records correctly mentioned pneumonia in the cause of death text but their corresponding ICD-10 codes failed to provide any code related to pneumonia, while 2 records were flagged because it included the sub-string "pneumonia" in the additional cause of death field. The death literal for these two records were "bacteremia due to Streptococcus pneumonia" and "Streptococcal Pneumoniae Septicemia", The remaining 13 errors were due to the entry of the death literals; in all cases the negation of 'aspiration pneumonia' either due to: 1) 'pneumonia' being in a separate cause of death field to 'aspiration' or 2) 'pneumonia' not being directly followed by 'aspiration' in the death text (example "pneumonia due to secondary aspiration"). A total of 20 false negatives were recorded, yielding a recall of 96%. The false negatives could be generalized into two categories: 1) misspellings of pneumonia on the death certificated (n = 8) and 2) appropriate pneumonia or influenza ICD-10 code was coded but the death literals did not mention an appropriate scanned phrase (n = 12). F-measure was also calculated at 96%. A high level of agreement was seen among keyword searching and the reference standard (kappa 0.95).

Death certificates pipeline

Utilizing the Death Certificates Pipeline (DCP), we identified 481 records as pneumonia and influenza deaths, 9 of which were false positives. The precision for this method was calculated at 98%. Like the keyword searching method, of the 9 false positives, 6 records mentioned pneumonia in the cause of death field but their corresponding ICD-10 codes failed to provide any code related to pneumonia and the remaining errors were due to the reporting of aspiration pneumonia on the death certificate. This method had only 1 false negative for the death literal stating "recurrent aspiration with pneumonia", thus yielding a recall at 99.8%, being less than keyword searching. F-measure was calculated at 99%. The level of agreement between the pipeline and the gold standard was almost perfect with a Cohen's kappa of 0.988.

The precision and recall scores that are reported above suggest that the DCP is a better method for identifying pneumonia and influenza deaths than keyword-searching. Therefore, we investigated if this observation is supported by statistical analysis. Performing a Fisher's exact test at $\alpha = 0.05$, significant difference was seen for both recall ($p = 1.742e\text{-}05$) and precision ($p = 0.026$). The McNemar's test result also showed DCP to be a better method with a p-value = 2.152e-05.

Analysis of failures

For the 472 pneumonia and influenza cases found by the reference standard, DCP correctly identified 471 cases, missed one case and incorrectly flagged nine cases. Most failures were due to discrepancies between the death literal and its respective ICD-10 code. For the only case which the pipeline did not match, the phrase 'recurrent aspiration with pneumonia' was present in the death literal. MetaMap coded this literal as aspiration pneumonia which was excluded from the CUI code list, but its respective ICD-10 included J189. For the 9 additional cases which were not present in the reference standard, we noticed two categories of errors: 1)cases where the string 'pneumonia' is present in the death literal but not coded into ICD-10 and 2) the reporting of aspiration pneumonia on the death certificate. The first category of errors was not due to MetaMap or the rule algorithm, but perhaps due to the coding process. As described earlier, MMDS produces entity axis and record axis codes. The entity axis codes would be a more appropriate reference standard for they provide the ICD 10 codes for the conditions or events reported as listed by the death certifier and maintains the order as written on the death certificate [43]; but as noted earlier only the record axis codes were made available for this study. The algorithm used to produce record axis codes from the entity axis data removes duplicate codes and contradictory diagnoses within the entity axis data to produce the more standardized record axis [44]. For example, if a medical examiner reports pneumonia with chronic obstructive pulmonary disease both conditions will be shown in entity axis code data. However, in record axis code data, they will be replaced with a single condition: Chronic obstructive pulmonary disease with acute lower respiratory infection (J44.0). We were unable to verify that codes related to pneumonia were present in the entity axis codes for the six cases; therefore, we can only speculate the reason for this failure.

The second category of errors was due to the reporting of aspiration pneumonia on the death certificate. In cases where the string "aspiration" and pneumonia" were not reported in the same text field MetaMap processed the string separately thus yielding two codes: one for aspiration and the other pneumonia, instead of one code for "aspiration pneumonia" [C0032290]. In an initial review of MetaMap we found MetaMap had difficulties processing the phrase "pneumonia secondary to acute aspiration", therefore, our rule detection algorithm excluded cases where the code for pneumonia and aspiration were present in the same text field.

Discussion

To our knowledge, this is the first published report on using a natural language processing tool and the UMLS to identify pneumonia and influenza deaths from death certificates. We found that automated coding and identification of pneumonia and influenza deaths is possible and computationally efficient. The Death Certificates Pipeline developed here was statistically different to keyword searching and has higher recall and precision when compared to the current semi-automatic methods in use by the CDC. A good recall is required to help capture the 'true' P-I deaths and a good precision is needed to avoidoverestimating the number of P-I deaths. This study also indicated that keyword searching underestimated pneumonia and influenza deaths in Utah. The simple keyword search method not only decreased recall and precision but also reduced the level of agreement. When reporting counts for surveillance purposes it's best to be as accurate as possible; however, there's a trade-off between recall and precision. For disease surveillance, increased precision enables public health officials to more accurately focus resources for control and prevention, therefore, although both methods had good precision the pipeline developed would be more advantageous to utilize.

MetaMap did an excellent job at extracting cause of deaths from free-form text which is consistent with the results of Reid et Al [25]. Most of the concepts were present in the UMLS which attributed good recall. Both recall and precision depended on the comprehensiveness of the CUI code list. The performance of this system is determined largely by the coverage of terms and sources in the UMLS. Both keyword searching and the system's weakest point is its lack of precision. Most of the concepts the system did not identify had either the aspiration text in another field or pneumonia was mentioned in the cause of death text but not coded (9 cases fit these criteria). The sample size was sufficient to show difference between the two methods. It is important to note that utilizing trained nosologists, who would manually code the death certificates, would have developed an absolute gold standard which may or may not be a better reference standard than ICD-10 codes. However, our motivation for utilizing ICD codes was influenced due to the fact that the use of ICD codes to identify all-cause pneumonia has been examined and has showed to be a valid tool for the identification of these cases [45,46].

In terms of timing, while keyword searching is faster than the DCP, our method is also sub 1/10 second range, which implies that it is possible to process the daily Utah deaths (~40) in approximately 5.47 seconds and all deaths in the US (~ 6646) in approximately 909.17 seconds using current hardware. This timing would be much faster than the minimum of two weeks to receive the coded data from the current CDC process. Moreover, these timings make it apparent that this system can be integrated in a real time surveillance system without introducing any additional bottlenecks.

There are several potential limitations with this analysis. First, the generalizability of the findings is limited because the death records were only from one institution. Although death certificates have a standardized format, the death registration process and the reviewing of death records differ by institutions. UDOH utilizes keyword searching to identify pneumonia and influenza cases, other institutions may use more accurate (manual review) or less accurate methods for finding cases. Second, a separate evaluation of the NLP component of the DCP was not performed. Further research is needed to examine the use of NLP on electronic death records across institutions and countries which may have different documentation procedures.

Conclusions and future work

This study shows that it is feasible to achieve high levels of accuracy when using NLP tools to identify cases of pneumonia and influenza cases from electronic death records while still providing a system that can be used for real time coding of death certificates. Identification of concept identifiers related to the CDC's case definition of pneumonia and influenza was very important in producing a highly accurate rule for the identification of these cases. Future work will aim to improve the preprocessing phase of the pipeline by providing the inclusion of the spellchecker used by the CDC's Mortality Medical Data System. Future work will also involve evaluating the flexibility (e.g. identification of different diseases) of the system to deploy the pipeline tool, along with other public health related analytical tools, as a grid service to provide to real time public health surveillance tool that uses data and services under the control of different administrative domains.

We have shown that it is feasible to automate the coding of electronic death records for real-time surveillance of deaths of public health concern. The performance of the Pipeline outperformed the performance of current methods, keyword searching, in the identification of pneumonia and influenza related deaths from death certificates. Therefore, the Pipeline has the potential to aid in the encoding of death certificates and is flexible to identify deaths due to other conditions of interest as the need arises.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All the authors contributed equally to this research. All authors read and approved the final manuscript.

Acknowledgments

This study has been supported in part by the grants from the National Library of Medicine (LM007124) and from the Centers of Disease Control and Prevention Center of Excellence (IP01HK000069-10).

Author details

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA. [2]Utah Department of Health, Salt Lake City, Utah, USA. [3]Center for High Performance Computing, University of Utah, Salt Lake City, Utah, USA.

Received: 22 September 2011 Accepted: 10 February 2012
Published: 8 May 2012

References

1. Mazick A, Participants of a workshop on mortality monitoring in Europe: **Monitoring excess mortality for public health action: potential for a future European network**. *Euro Surveill* 2007, **12**: E070104. Available from [http://www.eurosurveillance.org/ew/2007/070104.asp]
2. Declich S, Carter AO: **Public health surveillance: historical origins, methods and evaluation**. *Bull World Health Organ* 1994, **72**:285–304.
3. Galbraith NS: **Communicable disease surveillance.** In *Recent advances in community medicine, No 2*. Edited by Smith A. London, Churchill Livingstone; 1982:127–142.
4. Nogueira PJ, Machado A, Rodrigues E, Nunes B, Sousa L, Jacinto M, Ferreira A, Falcao JM, Ferrinho P: **The new automated daily mortality surveillance system in Portugal.** *Euro Surveill* 2010, **15**:pii 19529. Available from [http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19529]
5. Sartorius B, Jacobsen H, Törner A, Giesecke J: **Description of a new all cause mortality surveillance system in Sweden as a warning system using threshold detection algorithms**. *Eur J Epidemiol* 2006, **21**:181–9.
6. Simonsen L, Clarke MJ, Stroup DF, Williamson GD, Arden NH, Cox NJ: **A method for timely assessment of influenza-associated mortality in the United States**. *Epidemiology* 1997, **8**:390–5.
7. Haskey J: *Mortality surveillance 1968–1976, England and Wales. Deaths and rates by sex and age group for 8th revision causes, A-list and chapters.* London: Great Britain Office of Population Census and Surveys, Medical Statistics Division, Crown; 1978.
8. **Influenza WHO Fact sheet No. 211 revised March 2003**. [http://www.who.int/mediacentre/factsheets/fs211/en/4]
9. Fiore AE, Shay DK, Broder K, et al.: **Prevention and control of seasonal influenza with vaccines: recommendations of the Advisory Committee on Immunization Practices (ACIP),2009.** *MMWR Recomm Rep* 2009, **58**:1–52. [Erratum, *MMWR Recomm Rep* 2009, 58:896–7.]
10. Hall MJ, DeFrances CJ, Williams SN, Golosinskiy A, Schwartzman A: **National Hospital Discharge Survey: 2007 summary**. *Natl Health Stat Report* 2010, **26**:1–20.
11. Thompson WW, Shay DK, Weintraub E, *et al*: **Influenza-associated hospitalizations in the United States**. *JAMA* 2004, **292**:1333–40.
12. **American Lung Association State of Lung Disease in Diverse Communities 2010.** [http://www.lungusa.org/assets/documents/publications/lung-disease-data/solddc_2010.pdf]
13. Postma M, Bos JM, Van Gennep M, Jager JC, Baltussen R, Sprenger MJW: **Economic evaluation of influenza vaccination. Assessment for The Netherlands**. *Pharmacoeconomics* 1999, **16**(suppl1):33–40.
14. Thompson WW, Shay DK, Weintraub E, Brammer L, Cox N, Anderson LJ, Fukuda K: **Mortality associated with influenza and respiratory syncytial virus in the United States**. *JAMA* 2003, **289**:179–86.
15. **Estimating Seasonal Influenza-Associated Deaths in the United States: CDC Study Confirms Variability of Flu** [http://www.cdc.gov/flu/about/disease/us_flu-related_deaths.htm]
16. **Flu activity & surveillance: reports and surveillance methods in the United States** [http://www.cdc.gov/flu/weekly/fluactivity.htm]
17. **121 Cities Mortality Reporting System: history**. [http://www.cdc.gov/epo/dphsi/121hist.htm]
18. Wu WJ, Chaung JH: Real-time Surveillance of Pneumonia and Influenza Mortalities via the National Death Certificate System. [http://web.cdc.gov.tw/ct.asp?xItem=14106&ctNode=3842&mp=181]
19. Muscatello DJ, Morton PM, Evans I, Gilmour R: **Prospective surveillance of excess mortality due to influenza in New South Wales: feasibility and statistical approach**. *Commun Dis Intell* 2008, **32**:435–42.
20. Sager N: *Medical Language Processing: Computer Management of Narrative Data*. New York: Springer-Verlag; 1997.
21. **Death Certificate** [http://www.cdc.gov/NCIPC/pub-res/nvdrs-coding/Death-Certificate.pdf]
22. Anderson RN, Miniño AM, Hoyert DL, Rosenberg HM: **Comparability of cause of death between ICD-9 and ICD-10: preliminary estimates**. *Natl Vital Stat Rep* 2001, **49**:1–32.
23. World Health Organization: *International Statistical Classification of Disease and Related Health Problems, Tenth Revision Version for 2007 (ICD-10)*. 2006.
24. Harris, K: **Selected data editing procedures in an automated multiple cause of death coding system**. In *Proceedings of the Conference of European Statistics: 2–4 June 1999;Rome.*
25. Riedl B, Than N, Hogarth M: **Using the UMLS and Simple Statistical Methods to Semantically Categorize Causes of Death on Death Certificates.** In *AMIA Annual Symposium Proceeding 13 Nov* 2010; *Washington D.C.*2010:677–81.
26. Glenn D: **Description of the National Center for Health Statistics Software Systems and Demonstrations.** In *Proceedings of the international collaborative effort on automating mortality Volume I.July 1999.* Hyattsville, Maryland: National Center of Health Statistics; 1999.
27. **Toward an electronic death registration system in the United States: report of the Steering Committee to Reengineer the Death Registration Process**. *Am J Forensic Med Pathol* 1998, **19**:234–41.
28. **National Association for Public Health Statistics and Information Systems. Electronic Death Registration Systems by Jurisdiction Updated July 2011.** [http://www.naphsis.org/naphsis/files/ccLibraryFiles/Filename/000000001472/EDRS_Development_with_territories_July_2011.pdf]
29. Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, *et al*: **Classifying free-text triage chief complaints into syndromic categories with natural language processing.** *Artif Intell Med* 2005, **33**:31–40.
30. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ: **Automatic detection of acute bacterial pneumonia from chest X-ray reports.** *J Am Med Inform Assoc* 2000, **7**:593–604.
31. Friedman C, Shagina L, Lussier Y, Hripcsak G: **Automated encoding of clinical documents based on natural language processing.** *J Am Med Inform Assoc* 2004, **11**:392–402.
32. Gundlapalli AV, South BR, Chapman WW, Phansalkar S, Shen S, Delisle S, Perl Trish, Samore MH: **Using NLP on VA Electronic Medical Records to Facilitate Epidemiologic Case Investigations**. *Advances in Disease Surveillance* 2008, **5**:34.
33. Chapman WW, Dowling JN, Ivanov O, Gesteland PH, Espino JU, Wagner MM: **Evaluating natural language processing applications applied to outbreak and disease surveillance.** In *Proc 36th Symposium on the Interface: Computing Science and Statistics:26 May* 2004; *Baltimore.*
34. Aronson, A: **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.** In *Proc AMIA Symp*. 2001. 17–21.

35. Crowell J, Zeng Q, Ngo L, Lacroix EM: **A Frequency-based technique to improve the spelling suggestion rank in medical queries.** *J Am Med Inform Assoc* 2004, 11:179–185.
36. Browne AC, Divita G, Lu C, McCreedy L, Nace D: **TECHNICAL REPORT LHNCBC-TR-2003–003, Lexical Systems: A report to the Board of Scientific Counselors.** *Lister Hill National Center for Biomedical Communications, National Library of Medicine* 2003.
37. Browne AC, Divita G, Aronson AR, McCray AT: **UMLS language and vocabulary tools.** *Proceedings of the AMIA Annual Symposium. Washington DC, USA* 2003, 798.
38. Atkinson K: GNU Aspell. [http://aspell.sourceforge.net/]
39. Aronson, A.R: MetaMap: Mapping Text to the UMLS Metathesaurus. [http://skr.nlm.nih.gov/papers/references/metamap06.pdf]
40. Bodenreider O: **The Unified Language System (UMLS): integrating biomedical terminology.** *Nucleic Acids Res* 2004, 32:D267–70.
41. **UMLS distribution** [http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html]
42. R Core Development Team: *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing; 2009.
43. National Bureau of Economic Research: Entity axis codes [http://www.nber.org/mortality/1995/docs/entity95.txt]
44. National Center for Health Statistics: Documentation of the mortality tape file for 1997 data [http://wonder.cdc.gov/wonder/sci_data/mort/mcmort/type_txt/mcmort97.asp]
45. Whittle J, *et al*: **Community-acquired pneumonia: can it be defined with claims data?** *American Journal of Medical Quality* 1997, 12:187–193.
46. Skull SA, Andrews RM, Byrnes GB, Campbell DA, Nolan TM, Brown GV, Kelly HA: **ICD- 10 codes are a valid tool for identification of pneumonia in hospitalized patients aged > or = 65 years.** *Epidemiol Infect* 2008, 136:232–40. Epub 2007 Apr 20.

doi:10.1186/1472-6947-12-37
Cite this article as: Davis *et al.*: **Identification of pneumonia and influenza deaths using the death certificate pipeline.** *BMC Medical Informatics and Decision Making* 2012 **12**:37.

APPENDIX B

LIST OF ALGORITHMS IN THE

"SURVEILLANCE" PACKAGE

Table B.1: Algorithms available in the R "surveillance" package during summer 2012

Algorithm (Function)	Description
Cdc	Stroup et al. (1989)
Farrington	Farrington et al. (1996)
Farrington Flexible	An improved version of the Farrington algorithm
Rki	The system used at the Robert Koch Institute, Germany
Bayes	A Bayesian predictive posterior approach
Hmm	An online version of the Hidden Markov Model approach
Rogerson	An extended approach for surveillance for time varying Poisson means as documented in Rogerson and Yamada (2004)
Cusum	An approximate CUSUM method for time varying Poisson means as documented in Rossi et al (1999)
Glrnb	Likelihood and generalized likelihood ratio detectors for time varying Poisson and negative binomial distributed series documented in Höhle and Paul (2008)
Outbreak	Semiparametric surveillance of outbreaks by Frisén and Andersson (2009)
categoricalCUSUM	Includes change-point detection based on regression models for binomial and beta-binomial distributed response. Furthermore, multi-categorical models includes the multinomial logistic model, proportional odds model and the Bradley-Terry models
pairedbinCUSUM	Paired-binary approach taken in Steiner et al.
Algo.hhh	Held et al. (2005) and Paul et al. (2008)
Algo.twins	Held et al. (2006)
LRCUSUM.runlength	Markov Chain approximation for computing the run-length distribution
twinSIR	Continuous-time and discrete-space modeling as

	described in Höhle (2009). The appropriate data structure for this algorithm is found in the “epidata” class.
Twinstim	Continuous-time and continuous-space modeling as described in Meyer et al.
Stcd	Prospective space-time cluster detection

APPENDIX C

IMPLEMENTING PUBLIC HEALTH ANALYTICAL SERVICES: GRID ENABLING OF METAMAP

Citation: Davis K, Price RC, Facelli JC: **Implementing public health analytical services: Grid enabling of MetaMap**. In: *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on: 20 – 22 June 2013 2013. 113 – 118.*

Implementing Public Health Analytical Services: Grid Enabling of MetaMap

Kailah Davis, [a] Ronald C. Price [b] and Julio C. Facelli [*, b]
[a] Department of Biomedical Informatics, and [b] Center for High Performance Computing, University of Utah,
Salt Lake City, UT, USA
kailah.davisld),utah.edu. ron.price(a).utah.edu. iulio.facelli(a),utah.edu

Abstract

Public health data could be used to assist with public health surveillance and decision support. However, in most cases data has to be transformed into a coded format to make it computable and amiable to quasi real time analytical processing. Natural language processing (NLP) systems, which aim to accurately extract and encode biomedical information in a standard format, have a great potential in surveillance. NLP methods are complex, difficult, and expensive to implement. Its implementation, in most cases, is well beyond the technical expertise and resources available in Public Health organizations. Making NLP systems available as a service can greatly improve access to this methodology by public health officials and potentially enhance disease surveillance. MetaMap is a comprehensive biomedical NLP system, and has been shown to perform well for numerous applications. We describe how we have implemented MetaMap as a grid service to make it available to the public health community.

1. Introduction and Background

Public health surveillance and early detection of disease outbreaks have been a public health initiative for many centuries [1], Timely and accurate detection have been stressed because it can dramatically reduce the morbidity and mortality of a population [2], Historically, public health information systems supporting surveillance activities have been constructed to support specific program areas within health departments. These systems were often populated with data reported from healthcare providers, often via paper based forms, not surprisingly, leading to numerous "siloed" surveillance systems. Moreover, there's little effective communication and collaboration across all levels of public health; this has resulted in more than 300 separate biosurveillance efforts [3], The literature has indicated that many of these systems are disease specific, are not integrated or interoperable, and may be duplicative [4], In general the existing systems are not as comprehensive as needed and they do not use the most up to date computer technologies and data sources available today. New distributed approaches based on service oriented architectures are needed to construct the next generation of public health informatics systems.

Data useful for public health surveillance are usually available in unstructured, free-text format which can easily be read and understood by humans, but difficult for computers to decipher [5], The importance of unstructured or textual data for monitoring emerging infectious diseases and reporting public health outbreaks is growing. For example, the Public Health Agency of Canada's Global Public Health Intelligence Network (GPHIN) [6] continuously mines available free-text global news media sources and health and sciences websites in seven languages for disease outbreaks. Similarly, Project Argus [7] and ProMed-mail (the Program for Monitoring Emerging Diseases) [8] automatically monitor local and national media reports to rapidly disseminate information on outbreaks. HealthMap [9] and Global Health Monitor [10] also automatically collect news from the Internet about human and animal health, but they provide users a plot of the data on a World map using an established dictionary to map textual data and extract geographical location with mention of the disease. Indicators and warnings gathered from public domain reports of infectious disease outbreaks have been mined using keyword and text searching to create a heuristic staging model for the detection and assessment of outbreaks for both frontline personnel and decision-makers.

Despite the successes described above, the use of unstructured data is not widespread across public health [11] primarily because many of the current surveillance systems are unable to collect and analyze data from these non-traditional data sources [12]. Therefore, we should consider utilizing information that is collected by different public health systems.

Critical health information which can be used for surveillance purposes can be found in free-form text; however, the analyses of these data have inherent problems and should utilize structured data to successfully identify target population. It is also

978-l-4799-1053-3/13/$31.00 ©2013 IEEE

postulated that these data are ideal for surveillance systems because it provides greater sensitivity over specificity allowing greater detection of events of interest [5],The goal of natural language processing (NLP) is to classify, extract and encode free-text data. NLP has long been used to process clinical text and there have been many research groups developing NLP systems to aid in clinical research, decision support and quality assurance [13-15], For public health purposes, NLP has recently been applied to the domain disease surveillance for case identification [5, 16, 17] and to facilitate case investigation [14], with most of these application being focused on processing clinical data such as chief complaints [18]. These papers show the value of utilizing public health data for surveillance and argued that using NLP to acquire structured content is a crucial first step towards automated processing of public health data for surveillance purposes. However, NLP tools can quickly become difficult to implement due to reasons such as 1) the local deployment of complex systems can be costly and requires technical expertise generally uncommon at public health departments and 2) lack of training by public health officials in using complex command line parameters that are often required to extract meaningful structured data; thus making it difficult for non-technical users, such as an epidemiologist, to use during an outbreak investigation.

Grid architecture is a promising methodology to aggregate and analyze disparate, heterogeneous data and provide computational and analytic resources on demand for biomedical research [19]. Grid technology allows users to have access to services without knowledge of, or expertise with, or control of the infrastructure which supports them. Recently, grid technology has been embraced by the field of biomedical informatics [20] for research purposes particularly in imaging informatics and translational clinical research [21]. Grid computing applied to the public health domain has currently been on the rise and each of these study has demonstrated the potential benefits of leveraging grid technology in the public health domain [22-26]. Grid technology popularity is continually rising in the field of biomedical informatics because it encourages an ecosystem development culture [27], which has the potential to create virtual organizations for distributed fields such as public health.

A grid based version of an NLP tool would allow 1) deployment of a distributed service based architecture in a scalable grid environment with a well-defined service interface to NLP engines (thereby facilitating usage with a variety of end user tools); and 2) demonstrate the feasibility of enabling relevant applications to be a part of a national public health surveillance grid. This paper reports our efforts to make a natural language processing grid service and evaluate the effort needed to enable analytical grid services in order to create a public health virtual organization.

2. Related Work

Most NLP engines are designed for use in a centralized deployment and do not offer an interactive website or integrated service interface (MctaMap [28] has a Java API). Service oriented implementation of NLP engines have been focus of work in two previous projects. One project that has taken a service oriented approach to a distributed form of NLP is the Smntx [29], Smntx exposes a NLP engine through a standardized REST interface and facilitates semi-transparent parallel invocation of different NLP engines. Smntx allows users to execute real-time mining through complex queries and faceted search. The Smntx architecture focuses on cloud deployment. Another system, which is similar to Smntx, has been proposed by Carrel et al [30], but the authors have only reported preliminary work and a general description of the proposed architecture. Cancer Text Information Extraction System (caTIES) [31] also takes a service oriented approach but instead of cloud deployment this system is composed of a group of grid services that wraps the functionality and data sets. caTies is aimed at text mining information from surgical pathology reports using the National Cancer Institute Enterprise Vocabulary System and MetaMap. Although these systems are geared towards a distributed form of NLP their use can still be a non-trivial undertaking; both systems would require expertise to deploy and incorporate these systems into a public health grid.

3. Methods

MetaMap [28], developed by the National Library of Medicine (NLM), is useful in identifying biomedical concepts from free form textual input and maps them into concepts from the Unified Medical Language System (UMLS) Metathesaurus. MetaMap works by breaking the inputted text into words or phrases, map them to standard terms, and then match the terms to concepts in the Unified Medical Language System (UMLS) [32], For each matched phrase, MetaMap classifies it into a semantic type then returns the concept unique identifier (CUI) and the mapping options which are ranked according to the strength of the mapping. MetaMap was recently shown to be successful in the field of public health [18, 33]. However, like with many NLP tools, MetaMap typically run on local machine and consists of complex command line parameters thus limiting its use to technical users, who have access to local available data. Therefore, for this application to be used widely by public health users we envision grid enabling MetaMap and exposing its service to a public health grid environment. As noted, a grid enabled version of MetaMap can make a NLP engine available to wider range of users by providing client side web interface to facilitate intuitive non-expert use, allowing important processing capacity by using distributed data sources and incorporating the service into complex workflow analysis.

To our knowledge there has not been any attempt for grid enabling the MetaMap environment itself and using it as a part of a public health grid for epidemiologists. To demonstrate the usefulness of this tool to the public health surveillance we decided on the following scenario: An epidemiologist wants to perform interactive NLP text processing on death certificates to assess the severity of an influenza outbreak. On his local resources (client) he has available algorithms to identify disease ofinterests, but unable to properly analyze the unstructured data because he has no NLP tools available on his current machine or the substantial computational resources to quickly process the documents. The user can use a secure grid service to upload data to execute the text processing remotely, and transfer the results back to the client.

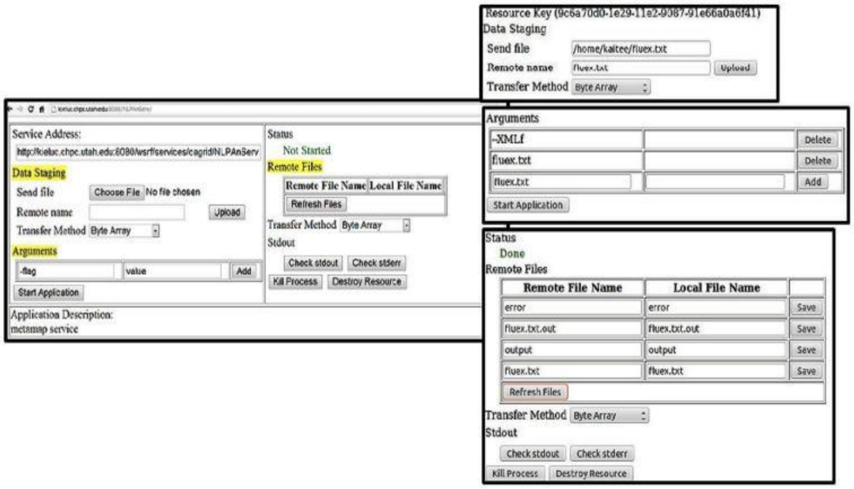
3.1 Creation and Validation

We accomplished our implementation using an iterative development approach with short iterations: iteration 1 involved installing and configuring MetaMap on a Linux based virtual machine and then wrapping MetaMap into a grid service; iteration 2 consisted of the deployment of the grid enabled MetaMap and, the final stage consisted of validating the service through proof of concept-a user case scenario.

To create the MetaMap grid service we used Introduce [34] and Grid Rapid Application Virtualization Interface (gRAVI) [35] which were developed by the caGrid project team. Introduce is an extensible toolkit which supports the development and deployment of WS/WSRF compliant grid services; this toolkit has many plug-ins to aid with development of different services. Grid Rapid Application Virtualization Interface (gRAVI) is an example of a plug-in that allows developers to quickly wrap and deploy applications as a Globus compliant grid service (http://dev.globus.org/wiki/Incubator/gRAVI). Through internal methods gRAVI allows a user to submit input files to the service and receive output over the grid.

The first step in our development required us to determine the parameter set required to execute MetaMap from the command line. We then used gRAVI and Introduce to wrap the MetaMap command line interface into a grid service. The service was then deployed to the caGrid training grid. After deployment, we invoked the service using the MetaMap grid service client that was also created automatically by Introduce and gRAVI.

Our second step was to use the grid service client to validate and illustrate the working principles of MetaMap in a grid environment. We realize that moving a technical application, like MetaMap, into a grid environment without providing a way for nontechnical users to use the application, particularly epidemiologists for the analysis of public health data, may not increase its usability because they are unfamiliar with the technical, programmatic and/or command line knowledge required to run the MetaMap software. As a result, although the service can be used in three interaction modes: the web interface, scientific workflows and programmatically, for the purpose of this project our validation was done through the web interface, which can be accessed at http://kieluc.chpc.utah.edu:8080/NLPAnServ/. The gRavi web client has three distinct sections: 1) data staging: used to upload files required to the working directory on the execution machine; 2) remote files: used to view all files in the working directory on the execution machine, and 3) arguments section: used button add to specify command-line arguments. We believe that the client servlet user interface would be the best option for demonstrating the service's usefulness to public health officials because, of the three options mentioned above, the web client provides a greater ease of use for the target user over programmatic interactions. Figure C.1 shows the webclient servlet; the three distinct sections mentionedabove are highlighted on the right.



Service Address: http://loeluc.chpc.utah.edu:8080/wsrf/services/cagrid/NLPAnServ

Data Staging

Send file: No file chosen

Transfer Method: Byte Array

Arguments: -flag | value

Status: Not Started

Remote Files

Remote File Name	Local File Name

Transfer Method: Byte Array

Stdout

Application Description:
metamap service

Resource Key (9c6a70d0-1e29-11e2-9087-91e66a0a6f41)

Data Staging

Send file: /home/kaltee/fluex.txt

Remote name: fluex.txt

Transfer Method: Byte Array

Arguments

-XMLf	
fluex.txt	
fluex.txt	

Status: Done

Remote Files

Remote File Name	Local File Name
error	error
fluex.txt.out	fluex.txt.out
output	output
fluex.txt	fluex.txt

Transfer Method: Byte Array

Stdout

Figure C.1: Using MetaMap Web Client

We used a subset of de-indentified death record data obtained by the Utah Department of Health to evaluate the performance of the grid enabled MetaMap. The records included a section describing the disease or condition directly leading to death, and any antecedent causes, co-morbid conditions and other significant contributing conditions. We randomly selected 6450 electronic death records all with multiple-cause-of-death fields (ex. underlying and immediate cause of death) from the Utah Department of Health (UDOH) for the period 1 January 2008 to 31 December 2008; these records resulted in 14314 death literals.

Using the grid web client, we uploaded a text a file of the death phrases for processing. The application then runs the text file on the execution server; this machine has all the necessary MetaMap services running to process the file and execute the specific scenario. Results of the analysis were returned to the user in the form of electronic files. The 14314 death literals were also processed on the local installation of MetaMap.

4. Results and Evaluation

The grid based version of MetaMap was successfully deployed and invoked. Using the the grid version of MetaMap gave exactly the same results as those as the executed in the stand-alone version. Moreover, the difference between execution times on the local installation (-0.169 seconds per death literals) and the grid version of MetaMap (0.132 seconds per death literal) was negligible. For this project we did not perform a rigorous analysis of the time and effort needed to implement MetaMap in a grid environment.

5. Discussion

As noted earlier, there has only been two service oriented approach to NLP, caTies and Smntx. However, to our knowledge there has not been so far any attempt for grid enabling the MetaMap environment itself and using it as user interface for accessing the Grid resources and invoking the Grid services. In this paper we have presented the integration of MetaMap in a grid environment. In the biomedical informatics community MetaMap is widely used and turned out as de facto standard. A Metamap grid enable version enables users to process documents in the grid and profit from the advantages of grid technology by provides a grid enabled environment for MetaMap. The presented use case showed the execution of a public health scenario and should be considered a proof of concept. Although the practical gains in terms of computational power still need to be done, we believe that the availability of grid enabled MetaMap will be of great use to epidemiologists.

Based on our experience we believe that it is possible to create a platform where public health departments contribute different services to the grid. However, when determining what services to deploy as a grid service developers should first determine what public health problem this service will address; and secondly, should create services that are used for both generalizable and specific to different public health functions.

6. Future Work

The work described is this paper is a summary of our experience of porting MetaMap to a grid environment. In order to fully take advantage of this work, future work will be aimed at developing a workflow that allows access that queries distributed public health death records, preprocess the records to be executed in MetaMap, process the document and output the document in an XML format. We intend to use Taverna to create this workflow where it will be deployed as a caGrid service.

7. Conclusion

We demonstrated that it is possible to easily deploy on a Grid applications for public health surveillance uses. We conclude that the techniques we used could be generalized to any application that has a command line interface. We believe that by providing a set of examples demonstrating the benefit of this technology to public health surveillance may lead to better, more collaborative system for public health surveillance.

9. Acknowledgements

This study has been supported in part by the grants from the National Library of Medicine (LM007124) and from the Centers of Disease Control and Prevention Center of Excellence (IP01HK000069-10).

References

[[1] S. Declich and A. O. Carter, "Public health surveillance: historical origins, methods and evaluation," Bulletin of the World Health Organization, vol. 72, p. 285, 1994.

[2] A. Zelicoff, J. Brillman, D. Forslund, J. George, S. Zink, S. Koenig, T. Staab, G. Simpson, E. Umland, and K. Bersell, "The Rapid Syndrome Validation Project (RSVP)," 2001, p. 771.

[3] N. B. A. Subcommittee, "Improving the Nation's Ability to Detect and Respond to 21st Century Urgent Health Threats: First Report of the National Biosurveillance Advisory Subcommittee," 2009.

[4] C. Staes, W. Xu, S. LeFevre, R. Price, S. Narus, A. Gundlapalli, R. Rolfs, B. Nangle, M. Samore, and J. Facelli, "A case for using grid architecture for state public health informatics: the Utah perspective," BMC Med Inform Decis Mak, vol. 9, p. 32, 2009.

[5] W. W. Chapman, "Natural Language Processing for Biosurveillance," Handbook of biosurveillance, p. 255, 2006.

[6] A. Mawudeku, R. Lemay, D. Werker, R. Andraghetti, and R. S. John, "The Global Public Health Intelligence Network," Infectious disease surveillance, pp. 304-317, 2007.

[7] H. Chen, P. Yan, D. Zeng, H. Chen, D. Zeng, and P. Yan, "Argus Infectious Disease Informatics." vol. 21, ed: Springer US, 2010, pp. 177-181.

[8] V. L. Yu and L. C. Madoff, "ProMED-mail: an early warning system for emerging diseases," Clinical Infectious Diseases, vol. 39, p. 227, 2004.

[9] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, "Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap Project," PLoS medicine, vol. 5, p. el51, 2008.

[10] S. Doan, Q. Hung-Ngo, A. Kawazoe, and N. Collier, "Global Health Monitor—a web-based system for detecting and mapping infectious diseases," 2008, pp. 951-956.

[11] "National Biosurveillance Strategy for Human Health Version 2," Centers for Disease Control and Prevention2010.

[12] O. United States. Government Accountability. (2005). Information technology federal agencies face challenges in implementing initiatives to improve public health infrastructure : report to congressional requesters Available: http://purl.access.gpo.gov/GPO/LPS617Q9

[13] W. W. Thompson, D. K. Shay, E. Weintraub, L. Brammer, C. B. Bridges, N. J. Cox, and K. Fukuda, "Influenza-associated hospitalizations in the United States," JAMA: the journal of the American Medical Association, vol. 292, pp. 1333-1340, 2004.

[14] A. V. Gundlapalli, B. R. South, W. W. Chapman, S. Phansalkar, S. Shen, S. Delisle, and S. M. H. Perl Trish, "Using NLP on VA Electronic Medical Records to Facilitate Epidemiologic Case Investigations,"

Advances in Disease Surveillance, vol. 5, p. 34, 2008.

[15] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak, "Automated encoding of clinical documents based on natural language processing," Journal of the American Medical Informatics Association, vol. 11, pp. 392-402, 2004.

[16] W. W. Chapman, J. N. Dowling, O. Ivanov, P. Gesteland, R. Olszewski, J. Espino, and M. M. Wagner, "Evaluating natural language processing applications applied to outbreak and disease surveillance," 2004.

[17] W. W. Chapman, A. V. Gundlapalli, B. R. South, and J. N. Dowling, "Natural language processing for biosurveillance," Infectious Disease Informatics and Biosurveillance, pp. 279-310,2011.

[18] K. Davis, C. Staes, R. Price, J. Duncan, S. Igo, and J. Facelli, "Real-time surveillance of influenza/pneumonia deaths: new strategies using grid computing and natural language processing," Automatically tracking diabetes using information in physicians' notes, p. 17, 2011.

[19] J. Facelli, "The Impact of Grid Computing in Biomedical Informatics," in INFOLAC2008-AAIM, Buenos Aires, Argentina, 2008.

[20] F. Berman, Grid computing : Making the global infrastructure a reality. Chichester: Wiley, 2003.

[21] A. Konagaya, "Trends in life science grid: from computing grid to knowledge grid," BMC bioinformatics, vol. 7, p. S10, 2006.

[22] V. Breton, A. L. da Costa, P. de Vlieger, L. Maigne, D. Sarramia, Y. M. Kim, D. Kim, H. Q. Nguyen, R. Reuillon, N. H. Truong, and Y. T. Wu, "Innovative in silico approaches to address avian flu using grid technology," Infect. Disord. Drug Targets Infectious Disorders - Drug Targets, vol. 9, pp. 358-365, 2009.

[23] P. De Vlieger, J. Y. Boire, V. Breton, Y. Legrd, J. Revillard, D. Sarramia, and L. Maigne, "Grid-enabled sentinel network for cancer surveillance," 2009.

[24] S. Hung, T. Hung, and J. Juang, "SARS Grid-An AG-Based Disease Management and Collaborative Platform," Studies in health technology and informatics, vol. 120, p. 217, 2006.

[25] F. A. B. da SILVA, H. F. Gagliardi, E. Gallo, M. A. Madope, V. C. Neto, I. T. Pisa, and D. Alves, "IntcgraEPI: a Grid-based epidemic surveillance system," Studies in health technology and informatics, vol. 126, p. 197, 2007.

[26] T. Savel, K. Hall, B. Lee, V. McMullin, M. Miles, J. Stinn, P. White, D. Washington, T. Boyd, and L. Lenert, "A Public Health Grid (PHGrid): Architecture and value proposition for 21st century public health," International Journal of Medical Informatics, vol. 79, pp. 523-529, 2010.

[27] J. C. Facelli, "An agenda for ultra-large-scale system research for global health informatics," ACM SIGHIT Record, vol. 2, pp. 12-12,2012.

[28] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MctaMap program," in Proceedings of the AMIA Symposium, 2001, p. 17.

[29] K. Chard, M. Russell, Y. A. Lussier, E. A. Mendonca, and J. C. Silverstein, "A cloud-based approach to medical NLP," AMIA Annu Symp Proc, vol. 2011, pp. 207-16, 2011.

[30] D. Carrell, "A strategy for deploying secure cloud-based natural language processing systems for applied research involving clinical text," in System Sciences (HICSS), 2011 44th Hawaii International Conference on, 2011, pp. 1-11.

[31] R. S. Crowley, M. Castine, K. Mitchell, G. Chavan, T. McSherry, and M. Feldman, "caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research," J Am Med Inform Assoc, vol. 17, pp. 253-64, May-Jim 2010.

[32] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," Nucleic Acids Res, vol. 32, pp. D267-D270, 2004.

[33] B. Riedl, N. Than, and M. Hogarth, "Using the UMLS and Simple Statistical Methods to Semantically Categorize Causes of Death on Death Certificates," 2010, p. 677.

[34] S. Hastings, S. Oster, S. Langella, D. Ervin, T. Kurc, and J. Saltz, "Introduce: an open source toolkit for rapid development of strongly typed grid services," Journal of Grid Computing, vol. 5, pp. 407-427, 2007.

[35] K. Chard, W. Tan, J. Boverhof, R. Madduri, and I. Foster, "Wrap scientific applications as WSRF grid services using gRAVI," 2009, pp. 83-90

APPENDIX D

A GRID BASED APPROACH TO SHARE PUBLIC HEALTH SURVEILLANCE APPLICATIONS: THE R EXAMPLE

Citation: Davis K, Facelli J: **A Grid Based Approach to Share Public Health Surveillance Applications: The R Example**. *Online Journal of Public Health Informatics* 2013, **5**(1).

ISDS 2012 Conference Abstracts

A Grid Based Approach to Share Public Health Surveillance Applications - The R Example

Kailah Davis*[1] and Julio Facelli[1, 2]

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA; [2]Center of High Performance Computing, University of Utah, Salt Lake City, UT, USA

Objective

This poster describes an approach which leverages grid technology for the epidemiological analysis of public health data. Through a virtual environment, users, particularly epidemiologists, and others unfamiliar with the application, can perform on-demand powerful statistical analyses.

Introduction

Currently, there's little effective communication and collaboration among public health departments. The lack of collaboration has resulted in more than 300 separate biosurveillance systems (1), which are disease specific, not integrated or interoperable, and may be duplicative (1). Grid architecture is a promising methodology to aid in building a decentralized health surveillance infrastructure because it encourages an ecosystem development culture (2), which has the potential to increase collaboration and decrease duplications.

Methods

This project had two major steps: creation and validation of the grid service. For the first step [creation of the service], we first determined the parameter set required to execute R from the command line. We then used the caGrid Introduce toolkit (3) and Grid Rapid Application Virtualization Interface (gRAVI) (4) to wrap the R command line interface into a grid service. The service was then deployed to the caGrid training grid. After deployment, the service was invoked using the R grid service client which was automatically created by Introduce and gRAVI.

Our second step was aimed at validating the service by using using the grid service client to illustrate the working principles of R in a grid environment. For this illustration, we selected the article by Hohle et al (5). In this article, the 'surveillance' package was developed to provide different algorithms for the detection of aberrations in routinely collected surveillance data. For validation purposes, only a subset of the analyses presented in the article, namely the Farrington and CUSUM algorithms, were reproduced. Using the grid web client, we uploaded the necessary data files for processing, as well as the Rscript which was used to replicate the results of (5). The application then ran the R script on the execution machine; this machine had all the necessary R packages needed for the specific scenario.

Results

The implementation of was validated by showing that the results of the original paper can be reproduced using gird based version of R. Figure C.1 shows the plots related to the steps described above; theplots illustrating the Farrington and CUSUM algorithms are seen tobe identical to that in (5).

Conclusions

We demonstrated that it is possible to easily deploy applications for public health surveillance uses. We conclude that the techniques we used could be generalized to any application that has a command line interface. Future work will be aimed creating a workflow to access data services and grid-enabled text processing and analytic tools. We believe that by providing a set of examples to demonstrate the benefit of this technology to public health surveillance infrastructure may provide insight that may lead to a better, more collaborative system of tools that will become the future of public health surveillance.

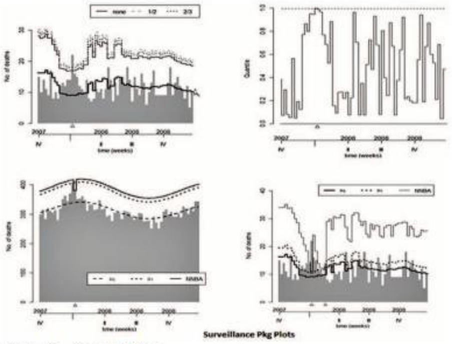


Figure D.1. Recreated Plots

Keywords

Grid computing; Public health grid; analytical service

Acknowledgments

This work was supported by NLM training grant #T15LM007124 and CDC Center of Excellence for Public Health Informatics # 1P01HK000069-10.

References

1. Subcommittee NBA. Improving the Nation's Ability to Detect and Respond to 21st Century Urgent Health Threats: First Report of the National Biosurveillance Advisory Subcommittee. 2009.
2. Facelli JC. An agenda for ultra-large-scale system research for global health informatics. ACM SIGHIT Record. 2012;2(1):12-.
3. Hastings S, Oster S, Langella S, Ervin D, Kurc T, Saltz J. Introduce: an open source toolkit for rapid development of strongly typed grid services. Journal of Grid Computing. 2007;5(4):407-27.
4. Chard K, Tan W, Boverhof J, Madduri R, Foster I, editors. Wrap scientific applications as WSRF grid services using gRAVI. 2009: IEEE.
5.Höhle M, Mazick A. Aberration detection in R illustrated by Danish mortality monitoring. Biosurveillance: Methods and Case Studies. 2010:215-37.

*Kailah Davis
E-mail: kailah.davis@utah.edu

REFERENCES

1. Declich S, Carter AO: **Public health surveillance: historical origins, methods and evaluation**. *Bulletin of the World Health Organization* 1994, **72**(2):285.

2. Zelicoff A, Brillman J, Forslund D, George J, Zink S, Koenig S, Staab T, Simpson G, Umland E, Bersell K: **The Rapid Syndrome Validation Project (RSVP)**. In*: 2001*. American Medical Informatics Association: 771.

3. Broome CV: **Overview, policy, and systems—Federal role in early detection preparedness systems**. In.: MMWR; 2005.

4. Keogh-Brown MR, Smith RD: **The economic impact of SARS: How does the reality match the predictions?** *Health policy* 2008, **88**(1):110-120.

5. APRIL O: **Outbreak of Swine-Origin Influenza A (H1N1) Virus Infection—Mexico, March-April 2009**. 2009.

6. Mair M, Maldin B, Smith B: **Passage of S. 3678: The Pandemic and All-Hazards Preparedness Act, UPMC**. *Center for Biosecurity, December* 2006, **20**.

7. Tokars J: **The BioSense Application**. In: *PHIN Conference Accessed on-line at http://0-www cdc gov mill1 sjlibrary org/biosense/files/Jerry Tokars ppt: 2006*.

8. **National Biosurveillance Strategy for Human Health Version 2**. In.: Centers for Disease Control and Prevention; 2010: 66.

9. **Biosurveillance nonfederal capabilities should be considered in creating a national biosurveillance strategy : report to congressional committees** [http://purl.fdlp.gov/GPO/gpo17798]

10. **Report to the President realizing the full potential of health information technology to improve healthcare for Americans the path forward** [http://purl.fdlp.gov/GPO/gpo2425]

11. Gesteland PH, Livnat Y, Galli N, Samore MH, Gundlapalli AV: **The EpiCanvas infectious disease weather map: an interactive visual exploration of temporal and spatial correlations**. *Journal of the American Medical Informatics Association* 2012, **19**(6):954-959.

12. Chapman WW: **Natural Language Processing for Biosurveillance**. *Handbook of biosurveillance* 2006:255.

13. Chapman WW, Gundlapalli AV, South BR, Dowling JN: **Natural language processing for biosurveillance**. *Infectious Disease Informatics and Biosurveillance* 2011:279-310.

14. **Information technology federal agencies face challenges in implementing initiatives to improve public health infrastructure : report to congressional requesters** [http://purl.access.gpo.gov/GPO/LPS61709]

15. Chapman WW, Dowling JN, Ivanov O, Gesteland P, Olszewski R, Espino J, Wagner MM: **Evaluating natural language processing applications applied to outbreak and disease surveillance**. In*: 2004*. Citeseer.

16. **Building a roadmap for health information systems interoperability for public health public health uses of electronic health record data : white paper** [http://static.ihe.net/Technical_Framework/upload/IHE-PHDSC_Public_Health_White_Paper_2008-07-29.pdf]

17. O'Carroll PW: **Public health informatics and information systems**. New York: Springer; 2003.

18. **6 U.S.C. Sec. 195b. National Biosurveillance Integration Center** [http://www.gpo.gov/fdsys/pkg/USCODE-2010-title6/pdf/USCODE-2010-title6-chap1-subchapIII-sec195b.pdf]

19. **Federal Health Architecture** [http://healthit.hhs.gov/portal/server.pt/community/%20healthit_hhs_gov__federal_health_architecture/1181]

20. **Public Health and Medical Preparedness** [http://www.fas.org/irp/offdocs/nspd/hspd-21.htm]

21. **Biodefense for the 21st Century** [http://www.fas.org/irp/offdocs/nspd/hspd-10.html]

22. **Pandemic and All-Hazards Preparedness Act** [http://www.fas.org/programs/bio/resource/documents/pl%20109-417.pdf]

23. **Public Law 110-53. Implementing Recomendations of the 9/11Commission Act of 2007** [http://intelligence.senate.gov/laws/pl11053.pdf]

24. Subcommittee NBA: **Improving the Nation's Ability to Detect and Respond to 21st Century Urgent Health Threats: First Report of the National Biosurveillance Advisory Subcommittee**. 2009.

25. Jenkins WO: **Biosurveillance: Efforts to Develop a National Biosurveillance Capability Need a National Strategy and a Designated Leader**: DIANE Publishing; 2010.

26. Dodaro GL: **Opportunities to Reduce Potential Duplication in Government Programs, Save Tax Dollars, and Enhance Revenue**. In.: DTIC Document; 2011.

27. Logsdon JK: **Biosurveillance technology: providing situational awareness through increased information sharing**. Monterey, California. Naval Postgraduate School; 2011.

28. Mawudeku A, Lemay R, Werker D, Andraghetti R, John RS: **The Global Public Health Intelligence Network**. *Infectious disease surveillance* 2007:304-317.

29. Chen H, Yan P, Zeng D, Chen H, Zeng D, Yan P: **Argus**

Infectious Disease Informatics. In., vol. 21: Springer US; 2010: 177-181.

30. Yu VL, Madoff LC: **ProMED-mail: an early warning system for emerging diseases**. *Clinical Infectious Diseases* 2004, **39**(2):227.

31. Brownstein JS, Freifeld CC, Reis BY, Mandl KD: **Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap Project**. *PLoS medicine* 2008, **5**(7):e151.

32. Doan S, Hung-Ngo Q, Kawazoe A, Collier N: **Global Health Monitor—a web-based system for detecting and mapping infectious diseases**. In*: 2008*. 951-956.

33. Systems CoEoNB, BioWatch, System tPH: **Biowatch and public health surveillance: evaluating systems for the early detection of biological threats**: Natl Academy Pr; 2010.

34. Bradley CA, Rolka H, Walker D, Loonsk J: **BioSense: implementation of a national early event detection and situational awareness system**. *MMWR Morb Mortal Wkly Rep* 2005, **54**:11-19.

35. Savel TG, Bronstein A, Duck W, Rhodes MB, Lee B, Stinn J, Worthen K, Deloitte Consulting L: **Using Secure Web Services to Visualize Poison Center Data for Nationwide Biosurveillance: A Case Study**. *Online Journal of Public Health Informatics* 2010, **2**(1).

36. Evans AS: **Surveillance and seroepidemiology**. In: *Viral Infections of Humans.* Springer; 1982: 43-64.

37. Pearl R: **Introduction to Medical Biometry and Statistics**. *Introduction to Medical Biometry and Statistics* 1930(Second Edition).

38. Johnson LR, Heymann DL: **Public Health Surveillance**. *Public health* 1997, **6**:02.01.

39. Graunt J: **Natural and Political Observations made upon the Bills of Mortality**: The Johns Hopkins Press; 1939.

40. Thacker SB, Berkelman RL: **Public health surveillance in the United States**. *Epidemiologic Reviews* 1988, **10**(1):164-190.

41. Eylenbosch W, Noah N: **Historical aspects**. *Surveillance in health and disease Oxford University Press, Oxford* 1988.

42. LANGMUIR AD: **William Farr: founder of modern concepts of surveillance**. *International Journal of Epidemiology* 1976, **5**(1):13-18.

43. Snow J: **On the mode of communication of cholera**: John Churchill; 1855.

44. Nogueira P, Machado A, Rodrigues E, Nunes B, Sousa L, Jacinto M, Ferreira A, Falcao J, Ferrinho P: **The new automated daily mortality surveillance system in Portugal**. 2010.

45. Sartorius B, Jacobsen H, Törner A, Giesecke J: **Description of a new all cause mortality surveillance system in Sweden as a warning system using threshold detection algorithms**. *European journal of epidemiology* 2006, **21**(3):181-189.

46. Simonsen L, Clarke MJ, Stroup DF, Williamson GD, Arden NH, Cox NJ: **A method for timely assessment of influenza-associated mortality in the United States**. *Epidemiology* 1997:390-395.

47. **Flu activity & surveillance: reports and surveillance methods in the United States** [http://www.cdc.gov/flu/weekly/fluactivity.htm]

48. **Leveraging Information Technology for Public Health Impact** [http://www.cdc.gov/od/ocio/]

49. Yasnoff WA, O Carroll PW, Koo D, Linkins RW, Kilbourne EM: **Public health informatics: improving and transforming public health in the information age**. *Journal of Public Health Management and Practice* 2000, **6**(6):67-75.

50. Dean A, Dean J, Burton A, Dicker R: **Epi Info: a general-purpose microcomputer program for public health information systems**. *American journal of preventive medicine* 1991, **7**(3):178.

51. Burton AH, Dean JA, Dean AG: **Software for data management and analysis in epidemiology**. *World health forum* 1990, **11**(1):75-77.

52. **National Electronic Telecommunications System for Surveillance--United States, 1990-1991**. *MMWR Morb Mortal Wkly Rep* 1991, **40**(29):502-503.

53. Friede A, Rosen DH, Reid JA: **CDC WONDER: a cooperative processing architecture for public health**. *Journal of the American Medical Informatics Association* 1994, **1**(4):303-312.

54. CDC: **Integrating public health information and surveillance systems: A report and recommendations from the CDC/ATSDR Steering Committee on Public Health Information and Surveillance System Development**. In. Edited by US Department of Health and Human Services PHS. Atlanta; GA; 1995.

55. Baker EL, Melton RJ, Stange PV, Fields ML, Koplan JP, Guerra FA, Satcher D: **Health reform and the health of the public. Forging community health partnerships**. *JAMA : the journal of the American Medical Association* 1994, **272**(16):1276-1282.

56. Baker EL, Friede A, Moulton AD, Ross DA: **CDC's Information Network for Public Health Officials (INPHO): a framework for integrated public health information and practice**. *Journal of public health management and practice : JPHMP* 1995, **1**(1):43-47.

57. **National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health and clinical medicine**. *Journal of public health management and practice : JPHMP* 2001, **7**(6):43-50.

58. Rotz LD, Koo D, O'Carroll PW, Kellogg RB, Sage MJ, Lillibridge SR: **Bioterrorism preparedness: planning for the future**. *Journal of public health management and practice : JPHMP* 2000, **6**(4):45-49.

59. Century IoMCoAtHotPits: **The Future of the Public's Health in the 21st Century**: National Academies Press; 2003.

60. Loonsk JW, McGarvey SR, Conn LA, Johnson J: **The Public Health Information Network (PHIN) Preparedness initiative**. *Journal of the American Medical Informatics Association : JAMIA* 2006, **13**(1):1-4.

61. Thacker SB, Qualters JR, Lee LM: **Public health surveillance in the United States: evolution and challenges**. *Morbidity and mortality weekly report Surveillance summaries (Washington, DC : 2002)* 2012, **61 Suppl**:3-9.

62. Tsui F-C, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM: **Technical description of RODS: a real-time public health surveillance system**. *Journal of the American Medical Informatics Association* 2003, **10**(5):399-408.

63. Chen H, Zeng D, Yan P: **Infectious Disease Informatics: Syndromic Surveillance for Public Health and Biodefense**, vol. 21: Springer Verlag; 2009.

64. Lombardo JS, Burkom H, Pavlin J: **ESSENCE II and the framework for evaluating syndromic surveillance systems**. *MMWR Morb Mortal Wkly Rep* 2004, **53**:159-165.

65. Gann R: **The BioSense - BioSurveillance to Improve Early Event Detection**. In: *SAS Global Forum: 2009; Washington D.C.*

66. Buehler JW, Isakov AP, Prietula MJ, Smith DJ, Whitney EA: **Preliminary Findings from the BioSense Evaluation Project**. *Advances in Disease Surveillance* 2007, **4**:237.

67. Kass-Hout TA, Massoudi B, Rojas-Smith L, Kaydos-Daniels S, Brownstein J, Buckeridge D, Buehler J: **BioSense program redesign**. In: *International Society for Disease Surveillance Conference 2010 Track 2: Public Health Surveillance: 2011*. 40.

68. Center HLS: **New information and intelligence needs in the 21st century threat environment**: Henry L. Stimson Center; 2008.

69. Carneiro HA, Mylonakis E: **Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks**. *Clinical Infectious Diseases* 2009, **49**(10):1557-1564.

70. Toner ES: **Creating situational awareness: A systems approach**. In: *Workshop on Medical Surge Capacity, Institute of Medicine Forum on Medical and Public Health Preparedness for Catastrophic Events: 2009*.

71. Patel M, Adighibe E, Lombardo J, Loschen W, Stewart M, Vernon MO: **Using Cloud Technology to Support Monitoring During High Profile Events**. *Online Journal of Public Health Informatics* 2013, **5**(1).

72. **National biosurveillance science and technology roadmap** [http://purl.fdlp.gov/GPO/gpo37707]

73. Friedman C, Hripcsak G: **Natural language processing and its future in medicine**. *Acad Med* 1999, **74**(8):890-895.

74. Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, Olszewski RT: **Classifying free-text triage chief complaints into syndromic categories with natural language processing**. *Artificial Intelligence in Medicine* 2005, **33**(1):31-40.

75. Spyns P: **Natural Language Processing**. *Methods of Information in Medicine* 1996, **35**(4):285-301.

76. Sager N, Friedman C, Lyman MS: **Medical language processing: computer management of narrative data**. 1987.

77. Sager N: **Syntactic analysis of natural language**. *Advances in Computers* 1967, **8**:153-188.

78. Sager N, Bross I, Story G, Bastedo P, Marsh E, Shedd D: **Automatic encoding of clinical narrative**. *Computers in Biology and Medicine* 1982, **12**(1):43-56.

79. Grishman R, Sager N, Raze C, Bookchin B: **The linguistic string parser**. In: *Proceedings of the June 4-8, 1973, national computer conference and exposition: 1973*. ACM: 427-434.

80. Sager N, Lyman M, Tick LJ, Borst F, Nhan NT, Revillard C, Su Y, Scherrer J-R: **Adapting a medical language processor from English to French**. In: *MEDINF089: proceedings of the Sixth International Conference on Medical Informatics Amsterdam, The Netherlands: Elsevier Science: 1989*. 548-553.

81. Oliver NC: **A sublanguage based medical language processing system for German**. 1992.

82. Lyman M, Sager N, Friedman C, Chi E: **Computer-Structured Narrative in Ambulatory Care: Its Use in Longitudinal Review of Clinical Data**. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care: 1985*. American Medical Informatics Association: 82.

83. Friedman C: **Medlee-a medical language extraction and encoding system**. *Columbia University, and Queens College of CUNY* 1995.

84. Jain NL, Knirsch CA, Friedman C, Hripcsak G: **Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports**. In: *Proceedings of the AMIA Annual Fall Symposium: 1996*. American Medical Informatics Association: 542.

85. Jain NL, Friedman C: **Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports**. In: *Proceedings of the AMIA Annual Fall Symposium: 1997*. American Medical Informatics Association: 829.

86. Friedman C, Knirsch C, Shagina L, Hripcsak G: **Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries**. In: *Proceedings of the AMIA Symposium: 1999*. American Medical Informatics Association: 256.

87. Mendonça EA, Haas J, Shagina L, Larson E, Friedman C: **Extracting information on pneumonia in infants using natural language processing of radiology reports**. *Journal of Biomedical Informatics* 2005, **38**(4):314-321.

88. Chiang J-H, Lin J-W, Yang C-W: **Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE)**. *Journal of the American Medical Informatics Association* 2010, **17**(3):245-252.

89. Lussier YA, Shagina L, Friedman C: **Automating icd-9-cm encoding using medical language processing: A feasibility study**. In: *Proceedings of the AMIA Symposium: 2000*. American Medical Informatics Association: 1072.

90. Lussier YA, Shagina L, Friedman C: **Automating SNOMED coding using medical language understanding: a feasibility study**. In: *Proceedings of the AMIA Symposium: 2001*. American Medical Informatics Association: 418.

91. !!! INVALID CITATION !!!

92. Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM: **Experience with a mixed semantic/syntactic parser**. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care: 1995*. American Medical Informatics Association: 284.

93. Haug PJ, Christensen L, Gundersen M, Clemons B, Koehler S, Bauer K: **A natural language parsing system for encoding admitting diagnoses**. In: *Proceedings of the AMIA Annual Fall Symposium: 1997*. American Medical Informatics Association: 814.

94. Fiszman M, Chapman WW, Evans SR, Haug PJ: **Automatic identification of pneumonia related concepts on chest x-ray reports**. In: *Proceedings of the AMIA Symposium: 1999*. American Medical Informatics Association: 67.

95. Trick WE, Chapman WW, Wisniewski MF, Peterson BJ, Solomon SL, Weinstein RA: **Electronic interpretation of chest radiograph reports to detect central venous catheters**. *Infection control and hospital epidemiology* 2003, **24**(12):950-954.

96. Denny JC, Smithers JD, Miller RA, Spickard A: **"Understanding" Medical School Curriculum Content Using KnowledgeMap**. *Journal of the American Medical Informatics Association* 2003, **10**(4):351-362.

97. Denny JC, Miller RA, Waitman LR, Arrieta MA, Peterson JF: **Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor**. *International Journal of Medical Informatics* 2009, **78**:S34-S42.

98. Doan S, Conway M, Phuong TM, Ohno-Machado L: **Natural Language Processing in Biomedicine: A Unified System Architecture Overview**. *arXiv preprint arXiv:14010569* 2014.

99. Aronson AR: **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program**. In: *Proceedings of the AMIA Symposium: 2001*. American Medical Informatics Association: 17.

100. Aronson AR: **Metamap: Mapping text to the umls metathesaurus**. *Bethesda, MD: NLM, NIH, DHHS* 2006.

101. Bodenreider O: **The unified medical language system (UMLS): integrating biomedical terminology**. *Nucleic acids research* 2004, **32**(suppl 1):D267-D270.

102. Gundlapalli AV, South BR, Chapman WW, Phansalkar S, Shen S, Delisle S, Perl Trish SMH: **Using NLP on VA Electronic Medical Records to Facilitate Epidemiologic Case Investigations**. *Advances in Disease Surveillance* 2008, **5**:34.

103. Gesteland PH, Gardner RM, Tsui F-C, Espino JU, Rolfs RT, James BC, Chapman WW, Moore AW, Wagner MM: **Automated syndromic surveillance for the 2002 Winter Olympics**. *Journal of the American Medical Informatics Association* 2003, **10**(6):547-554.

104. Wagner MM, Moore AW, Aryel RM: **Handbook of biosurveillance**: Academic Press; 2006.

105. Hogan WR, Wagner MM: **Sales of over-the-counter healthcare products**. *Handbook of Biosurveillance Amsterdam* 2006.

106. Riedl B, Than N, Hogarth M: **Using the UMLS and Simple Statistical Methods to Semantically Categorize Causes of Death on Death Certificates**. In*: 2010*. American Medical Informatics Association: 677.

107. Foster I, Kesselman C, Tuecke S: **The anatomy of the grid: Enabling scalable virtual organizations**. *International Journal of High Performance Computing Applications* 2001, **15**(3):200-222.

108. Berman F, Fox GC, Hey AJG: **Grid computing : making the global infrastructure a reality**. In*: 2003; New York*. J. Wiley.

109. Berman F: **Grid computing : Making the global infrastructure a reality**. Chichester: Wiley; 2003.

110. **The caBig Impact** [http://cabig.cancer.gov/perspectives/impact/]

111. Breton V, Dean K, Solomonides T, Blanquer I, Hernandez V, Medico E, Maglaveras N, Benkner S, Lonsdale G, Lloyd S: **The healthgrid white paper**. *Studies in health technology and informatics* 2005, **112**(249-321):28-39.

112. Erberich SG, Silverstein JC, Chervenak A, Schuler R, Nelson MD, Kesselman C: **Globus MEDICUS - federation of DICOM medical imaging devices into healthcare Grids**. *Stud Health Technol Inform* 2007, **126**:269-278.

113. Sharma A, Pan T, Cambazoglu BB, Gurcan M, Kurc T, Saltz J: **VirtualPACS—a federating gateway to access remote image data resources over the grid**. *Journal of Digital Imaging* 2009, **22**(1):1-10.

114. Bertero M, Bonetto P, Carracciuolo L, D'Amore L, Formiconi A, Guarracino MR, Laccetti G, Murli A, Oliva G: **MedIGrid: a Medical Imaging application for computational Grids**. In: *Parallel and Distributed Processing Symposium, 2003 Proceedings International: 2003*. IEEE: 8 pp.

115. Bagarinao E, Matsuo K, Tanaka Y, Sarmenta LF, Nakai T: **Enabling on-demand real-time functional MRI analysis using grid technology**. *Methods Inf Med* 2005, **44**(5):665-673.

116. Khalil I, Fahim S: **CardioGrid: ECG analysis on demand to detect cardiovascular abnormalities**. In: *Information Technology and Applications in Biomedicine, 2009 ITAB 2009 9th International Conference on: 4-7 Nov. 2009 2009*. 1-5.

117. Pan TC, Gurcan MN, Langella SA, Oster SW, Hastings SL, Sharma A, Rutt BG, Ervin DW, Kurc TM, Siddiqui KM *et al*: **Informatics in radiology: GridCAD: grid-based computer-aided detection system**. *Radiographics : a review publication of the Radiological Society of North America, Inc* 2007, **27**(3):889-897.

118. Estrella F, McClatchey R, Rogulin D: **The MammoGrid Virtual Organisation - Federating Distributed Mammograms**. *Stud Health Technol Inform* 2005, **116**:935-940.

119. Spezi E, Lewis G: **An overview of Monte Carlo treatment planning for radiotherapy**. *Radiation protection dosimetry* 2008, **131**(1):123-129.

120. Jacq N, Salzemann J, Jacq F, Legré Y, Medernach E, Montagnat J, Maaß A, Reichstadt M, Schwichtenberg H, Sridhar M: **Grid-enabled virtual screening against malaria**. *Journal of Grid Computing* 2008, **6**(1):29-43.

121. Lee HC, Salzemann J, Jacq N, Chen HY, Ho LY, Merelli I, Milanesi L, Breton V, Lin SC, Wu YT *et al*: **Grid-enabled high-throughput in silico screening against influenza A neuraminidase**. *IEEE Trans Nanobioscience* 2006, **5**(4):288-295.

122. Krishnan A: **GridBLAST: a Globus ‐ based high ‐ throughput implementation of BLAST in a Grid computing framework**. *Concurrency and Computation: Practice and Experience* 2005, **17**(13):1607-1623.

123. Konishi F, Konagaya A: **The architectural design of high-throughput BLAST services on OBIGrid**. In: *Grid Computing in Life Science.* Springer; 2005: 32-42.

124. Afgan E, Bangalore P: **Dynamic BLAST–a Grid Enabled BLAST**. *IJCSNS International Journal of Computer Science and Network Security* 2009, **9**(4).

125. Carvalho PC, Glória RV, de Miranda AB, Degrave WM: **Squid–a simple bioinformatics grid**. *BMC bioinformatics* 2005, **6**(1):197.

126. Sinnott R, Ajayi O, Jiang J, Stell A, Watt J: **User oriented access to secure biomedical resources through the grid**. 2006.

127. Bayer M, Campbell A, Virdee D: **A GT3 based BLAST grid service for biomedical research**. In: *Proceedings of the UK e-Science All Hands Meeting: 2004*. Citeseer: 1019-1023.

128. Sulakhe D, Rodriguez A, D'Souza M, Wilde M, Nefedova V, Foster I, Maltsev N: **GNARE: automated system for high-throughput genome analysis with grid computational backend**. *Journal of clinical monitoring and computing* 2005, **19**(4):361-369.

129. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A: **Taverna: a tool for the composition and enactment of bioinformatics workflows**. *Bioinformatics* 2004, **20**(17):3045-3054.

130. Taylor I, Shields M, Wang I, Harrison A: **Visual grid workflow in Triana**. *Journal of Grid Computing* 2005, **3**(3-4):153-169.

131. Taylor I, Shields M, Wang I, Harrison A: **The triana workflow environment: Architecture and applications**. In: *Workflows for e-Science.* Springer; 2007: 320-339.

132. Abouelhoda M, Issa SA, Ghanem M: **Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support**. *BMC bioinformatics* 2012, **13**(1):77.

133. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J: **Galaxy: a platform for interactive large-scale genome analysis**. *Genome research* 2005, **15**(10):1451-1455.

134. Ghanem M, Curcin V, Wendel P, Guo Y: **Building and using analytical workflows in discovery net**. *Data Mining Techniques in Grid Computing Environments, Wiley-Blackwell* 2008:119-240.

135. Deelman E, Blythe J, Gil Y, Kesselman C, Mehta G, Patil S, Su M-H, Vahi K, Livny M: **Pegasus: Mapping scientific workflows onto the grid**. In: *Grid Computing: 2004*. Springer: 11-20.

136. Bradley J, Brown C, Carpenter B, Chang V, Crisp J, Crouch S, De Roure D, Newhouse S, Li G, Papay J: **The OMII software distribution**. 2006.

137. Van Der Aalst WM, Ter Hofstede AH: **YAWL: yet another workflow language**. *Information systems* 2005, **30**(4):245-275.

138. Altintas I, Berkley C, Jaeger E, Jones M, Ludascher B, Mock S: **Kepler: An extensible system for design and execution of scientific workflows**. In*: 2004*. Ieee: 423-424.

139. Linke B, Giegerich R, Goesmann A: **Conveyor: a workflow engine for bioinformatic analyses**. *Bioinformatics* 2011, **27**(7):903-911.

140. Shah SP, He DY, Sawkins JN, Druce JC, Quon G, Lett D, Zheng GX, Xu T, Ouellette BF: **Pegasys: software for executing and integrating analyses of biological sequences**. *BMC bioinformatics* 2004, **5**(1):40.

141. Tiwari A, Sekhar AK: **Review Article: Workflow based framework for life science informatics**. *Computational Biology and Chemistry* 2007, **31**(5-6):305-319.

142. Curcin V, Ghanem M: **Scientific workflow systems-can one size fit all?** In: *Biomedical Engineering Conference, 2008 CIBEC 2008 Cairo International: 2008*. IEEE: 1-9.

143. Breton V, da Costa AL, de Vlieger P, Maigne L, Sarramia D, Kim YM, Kim D, Nguyen HQ, Reuillon R, Truong NH *et al*: **Innovative in silico approaches to address avian flu using grid technology**. *Infect Disord Drug Targets Infectious Disorders - Drug Targets* 2009, **9**(3):358-365.

144. De Vlieger P, Boire JY, Breton V, Legré Y, Revillard J, Sarramia D, Maigne L: **Grid-enabled sentinel network for cancer surveillance**. 2009.

145. Hung S, Hung T, Juang J: **SARS Grid-An AG-Based Disease Management and Collaborative Platform**. *Studies in health technology and informatics* 2006, **120**:217.

146. da SILVA FAB, Gagliardi HF, Gallo E, Madope MA, Neto VC, Pisa IT, Alves D: **IntegraEPI: a Grid-based epidemic surveillance system**. *Studies in health technology and informatics* 2007, **126**:197.

147. Boyd T, Lee B, Savel T, Stinn J: **An example of the use of Public Health Grid (PHGrid) technology during the 2009 H1N1 influenza pandemic**. *International Journal of Grid and Utility Computing* 2011, **2**(2):148-155.

148. Staes CJ, Xu W, LeFevre SD, Price RC, Narus SP, Gundlapalli A, Rolfs R, Nangle B, Samore M, Facelli JC: **A case for using grid architecture for state public health informatics: the Utah perspective**. *BMC medical informatics and decision making* 2009, **9**(1):32.

149. Easterby-Smith M, Thorpe R, Jackson P: **Management research**: Sage Publications; 2012.

150. Friedman CP, Wyatt J: **Evaluation methods in biomedical informatics**: Springer; 2006.

151. Bernard HR, Bernard HR: **Social research methods: Qualitative and quantitative approaches**: Sage; 2012.

152. De Villiers M: **Interpretive research models for Informatics: action research, grounded theory, and the family of design-and development research**. *Alternation* 2005, **12**(2):10-52.

153. Heathfield H, Peel V, Hudson P, Kay S, Mackay L, Marley T, Nicholson L, Roberts R, Williams J: **Evaluating large scale health information systems: from practice towards theory**. In: *Proceedings of the AMIA Annual Fall Symposium: 1997*. American Medical Informatics Association: 116.

154. **Research Methods Infosystems**. In.

155. Lores: **Exploratory, Descriptive, and Causal Research Designs**. In.; 2012.

156. van Teijlingen E, Hundley V: **The importance of pilot studies**. *Social research update* 2001(35):1-4.

157. Struwig M, Struwig F, Stead G: **Planning, Reporting & Designing Research**: Pearson South Africa; 2001.

158. Brewer J, Hunter A: **Multimethod research: A synthesis of styles**: Sage Publications, Inc; 1989.

159. Israel RA: **Automation of mortality data coding and processing in the United States of America**. *World health statistics quarterly Rapport trimestriel de statistiques sanitaires mondiales* 1989, **43**(4):259-262.

160. Harris K: **Selected data editing procedures in an automated multiple cause of death coding system**. In: *Proc Conference Eur Stat: 1999*.

161. Organization WH: **International Statistical Classification of Disease and Related Health Problems, Tenth Revision Version for 2007 (ICD-10)**. In.; 2006.

162. Browne AC, Divita G, Aronson AR, McCray AT: **UMLS language and vocabulary tools**. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium* 2003:798.

163. Divita G, Lu C, McCreedy L, Nace D: **Lexical Systems; A report to the Board of Scientific Counselors September 2003 Allen C. Browne**. 2003.

164. Crowell J, Zeng Q, Ngo L, Lacroix E-M: **A frequency-based technique to improve the spelling suggestion rank in medical queries**. *Journal of the American Medical Informatics Association* 2004, **11**(3):179-185.

165. Atkinson K: **Gnu aspell**. In.: Online: http://aspell. net/(seen August 13, 2011); 2005.

166. Team RC: **R: A language and environment for statistical computing**. In.: ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. url: http://www. R-project. org; 2005.

167. Davis K, Staes C, Duncan J, Igo S, Facelli JC: **Identification of pneumonia and influenza deaths using the death certificate pipeline**. *BMC medical informatics and decision making* 2012, **12**(1):37.

168. Ihaka R, Gentleman R: **R: A language for data analysis and graphics**. *Journal of computational and graphical statistics* 1996:299-314.

169. Gómez-Rubio V, Ferrándiz-Ferragud J, López-Quílez A: **Detecting clusters of disease with R**. *Journal of geographical systems* 2005, **7**(2):189-206.

170. Kim AY, Wakefield J: **R Data and Methods for Spatial Epidemiology: the SpatialEpi Package**. 2010.

171. Aragon T: **Epitools: R Package for Epidemiologic Data and Graphics**. In.; 2007.

172. Höhle M: **: An R package for the monitoring of infectious diseases**. *Computational Statistics* 2007, **22**(4):571-582.

173. Stroup DF, Williamson GD, Herndon JL, Karon JM: **Detection of aberrations in the occurrence of notifiable diseases surveillance data**. *Stat Med* 1989, **8**(3):323-329.

174. Farrington C, Andrews N, Beale A, Catchpole M: **A statistical algorithm for the early detection of outbreaks of infectious disease**. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1996:547-563.

175. Rossi G, Lampugnani L, Marchi M: **An approximate CUSUM procedure for surveillance of health events**. *Stat Med* 1999, **18**(16):2111-2122.

176. Rogerson PA, Yamada I: **Approaches to syndromic surveillance when data consist of small regional counts**. *Morbidity and Mortality Weekly Report* 2004, **53**(Suppl.):79-85.

177. Höhle M, Paul M: **Count data regression charts for the monitoring of surveillance time series**. *Computational Statistics & Data Analysis* 2008, **52**(9):4357-4368.

178. Held L, Höhle M, Hofmann M: **A statistical framework for the analysis of multivariate infectious disease surveillance counts**. *Statistical Modelling* 2005, **5**(3):187-199.

179. Paul M, Held L, Toschke A: **Multivariate modelling of infectious disease surveillance data**. *Stat Med* 2008, **27**(29):6250-6267.

180. Le Strat Y, Carrat F: **Monitoring epidemiologic surveillance data using hidden Markov models**. *Stat Med* 1999, **18**(24):3463-3478.

181. Held L, Hofmann M, Höhle M, Schmid V: **A two-component model for counts of infectious diseases**. *Biostatistics* 2006, **7**(3):422-437.

182. Fenstermacher D, Street C, McSherry T, Nayak V, Overby C, Feldman M: **The Cancer Biomedical Informatics Grid (caBIG(tm))**. *Conf Proc IEEE Eng Med Biol Soc* 2005, **1**:743-746.

183. Saltz J, Oster S, Hastings S, Langella S, Kurc T, Sanchez W, Kher M, Manisundaram A, Shanbhag K, Covitz P: **caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid**. *Bioinformatics* 2006, **22**(15):1910-1916.

184. Rumbaugh J, Jacobson I, Booch G: **The unified modeling language reference manual**. 1999.

185. Architect E: **Sparx systems**. In.: Inc; 2011.

186. Phillips J, Chilukuri R, Fragoso G, Warzel D, Covitz PA: **The caCORE Software Development Kit: Streamlining construction of interoperable biomedical information services**. *BMC Medical Informatics and Decision Making* 2006, **6**(1):2.

187. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, Coronado Sd, Reeves DM, Hadfield JB, Ludet C: **caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability**. *Journal of biomedical informatics* 2008, **41**(1):106-123.

188. Hastings S, Oster S, Langella S, Ervin D, Kurc T, Saltz J: **Introduce: an open source toolkit for rapid development of strongly typed grid services**. *Journal of Grid Computing* 2007, **5**(4):407-427.

189. Chard K, Tan W, Boverhof J, Madduri R, Foster I: **Wrap scientific applications as WSRF grid services using gRAVI**. In*: 2009*. IEEE: 83-90.

190. Davis K, Price RC, Facelli JC: **Implementing public health analytical services: Grid enabling of MetaMap**. In: *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on: 20-22 June 2013 2013*. 113-118.

191. Davis K, Facelli J: **A Grid Based Approach to Share Public Health Surveillance Applications-The R Example**. *Online Journal of Public Health Informatics* 2013, **5**(1).

192. Fleiss JL: **Statistical methods for rates and proportions**. In.: Wiley (New York); 1981.

193. **National Bureau of Economic Research: Entity axis codes.** [http://www.nber.org/mortality/1995/docs/entity95.txt]

194. Statistics NCfH: **Documentation of the mortality tape file for 1997 data**. In. Edited by Statistics NCfH.

195. Höhle M, Mazick A: **Aberration detection in R illustrated by Danish mortality monitoring**. *Biosurveillance: Methods and Case Studies* 2010:215-237.

196. Boyd T, Savel T, Kesarinath G, Lee B, Stinn J: **The use of public health grid technology in the united states centers for disease control and prevention h1n1 pandemic response**. In: *Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on: 2010*. IEEE: 974-978.