Lecture Notes in Bioinformatics 5354

Subseries of Lecture Notes in Computer Science

Daniel Zeng Hsinchun Chen Henry Rolka
Bill Lober (Eds.)

Biosurveillance and Biosecurity

International Workshop, BioSecure 2008
Raleigh, NC, USA, December 2, 2008
Proceedings

Springer

Volume Editors

Daniel Zeng
MIS Department, University of Arizona
Tucson, AZ, USA
and Chinese Academy of Sciences
E-mail: zeng@email.arizona.edu

Hsinchun Chen
MIS Department, University of Arizona
Tucson, AZ, USA
E-mail: hchen@eller.arizona.edu

Henry Rolka
US CDC, National Center for Public Health Informatics
Atlanta, GA, USA
E-mail: hrr2@cdc.gov

Bill Lober
Health Sciences Building, University of Washington
Seattle, WA, USA
E-mail: lober@u.washington.edu

Library of Congress Control Number: Applied for

CR Subject Classification (1998): J.3, H.2, H.5

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-89745-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-89745-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12577477 06/3180 5 4 3 2 1 0

Preface

The 2008 Biosurveillance and Biosecurity Workshop (BioSecure 2008) was built on the success of the two U.S. National Science Foundation-sponsored Biosurveillance Workshops. The inaugural 2006 workshop was hosted by the University of Arizona's NSF BioPortal Center. It attracted more than 35 participants from academic institutions, industry, and public health agencies, and achieved its objective of bringing together infectious disease informatics (IDI) researchers and practitioners to discuss selected topics directly relevant to data sharing and analysis for real-time animal and public health surveillance. The 2007 meeting was held in New Brunswick, New Jersey, co-located with the 2007 IEEE International Conference on Intelligence and Security Informatics, and met with tremendous success. Researchers from a wide range of backgrounds, including biosecurity, epidemiology, statistics, applied mathematics, information systems, computer science and machine learning/data mining, contributed formal papers to the workshop and actively participated in the meeting along with practitioners from both government agencies and industry. More than 65 people attended the one-day workshop, representing major research labs across multiple disciplines, key industry players, and a range of government entities.

BioSecure 2008 continued this workshop series aiming to achieve the following objectives: (a) review and examine various informatics approaches for health surveillance and biosecurity from both technological and policy perspectives; and (b) discuss and compare various systems approaches and algorithms of relevance to biosurveillance and biosecurity. The specific emphasis of the 2008 meeting was to encourage information and computer science (including informatics, statistics, modeling and decision sciences, data management, and IT) researchers to join the public health surveillance and biosecurity community to conduct high-impact and innovative research.

We are pleased to have received many outstanding contributions from IDI research groups and practitioners from around the world. The one-day program included one invited presentation, 18 papers, and an all-inclusive poster session.

BioSecure 2008 was jointly hosted by the University of Arizona, the U.S. Centers for Disease Control and Prevention, and the University of Washington. We wish to express our gratitude to all workshop Program Committee members, who provided high-quality, timely, valuable and constructive review comments. In addition, we sincerely appreciate the efforts by our Government Liaisons, Sylvia Spengler from U.S. National Science Foundation, Donald Noah, U.S. Department of Homeland Security, and Daniel M. Sosin, U.S. Centers for Disease Control and Prevention, to broaden workshop participation from related academic and practitioner communities. We also would like to thank Catherine A. Larson, and several members of the Artificial Intelligence Laboratory and the Intelligent Systems and Decisions Laboratory at the University of Arizona for their excellent support.

BioSecure 2008 was held as part of the pre-conference workshop series at the International Society for Disease Surveillance (ISDS) Seventh Annual Conference. We wish to thank the ISDS society officers, meeting organizers, and support staff for their cooperation and assistance. We also wish to acknowledge the Springer LNCS editorial and production staff for their professionalism and continued support for intelligence and security informatics, IDI, and related events. Our sincere gratitude goes to the U.S. National Science Foundation as the main sponsor.

Daniel Zeng
Hsinchun Chen
Henry Rolka
Bill Lober

Organization

Organizing Committee

Conference Co-chairs

Hsinchun Chen	University of Arizona
Henry Rolka	U.S. Centers for Disease Control and Prevention (CDC)
Daniel Zeng	University of Arizona and Chinese Academy of Sciences
Bill Lober	University of Washington

Government Liaisons

Sylvia Spengler	U.S. National Science Foundation
Donald Noah	U.S. Department of Homeland Security
Daniel M. Sosin	U.S. Centers for Disease Control and Prevention

Program Committee

David Banks	Duke University
John Berezowski	Alberta (Canada) Agriculture
Ian Brooks	University of Illinois at Urbana-Champaign
David L. Buckeridge	McGill University
Howard Burkom	Johns Hopkins University
Jean-Paul Chretien	Walter Reed Army Institute of Research and Johns Hopkins University
Greg Cooper	University of Pittsburgh
Dan Desmond	SIMI Group
Daniel Ford	IBM
Ron Fricker	Naval Postgraduate School
Colin Goodall	AT&T
Ivan Gotham	New York State Dept. of Health
Valerie Gregg	University of Albany
Paul Hu	University of Utah
Xiaohua (Tony) Hu	Drexel University
Jesse Huang, Peking	Union Medical College
C.-C. King	National Taiwan University
Ken Kleinman	Harvard University
Ken Komatsu	Arizona Dept. of Health Services
Eileen Koski	Quest Diagnostics Incorporated
Sheri Lewis	Johns Hopkins University

Cecil Lynch	Ontoreason
Colleen Martin,	U.S. CDC
Jean O'Connor	U.S. CDC
Theresa Pardo	University of Albany
Michelle Podgornik	U.S. DHS & CDC
Ben Reis	Massachusetts Institute of Technology
Debbie Travers	University of North Carolina at Chapel-Hill
David Walker	U.S. CDC
Quanyi Wang	Beijing CDC
Xiaohui Zhang	Scientific Technologies Corp

Table of Contents

Informatics Infrastructure and Policy Considerations

Network-Based Data Analytics

Biosurveillance Models and Outbreak Detection

Model Assessment and Case Studies

Environmental Biosurveillance and Case Studies

Public Health Information Fusion for Situation Awareness

Henry Rolka, Jean C. O'Connor, and David Walker

Office of Critical Information Integration and Exchange, National Center for Zoonotic, Vector-Borne and Enteric Diseases, Centers for Disease Control and Prevention, Atlanta, GA
{Henry Rolka,Jean O'Connor,David Walker,LNCS}@Springer.com

Abstract. Recent events, including the terrorist attacks in the fall of 2001, the spread of Severe Acute Respiratory Syndrome (SARS), and Hurricane Katrina, highlight the need for real-time information exchange to enhance government's awareness and understanding of public health events in order to detect and respond as those events unfold. This paper describes the planned approach of the Centers for Disease Control and Prevention (CDC)'s Office of Critical Information Integration and Exchange (OCIIX) in meeting that need through the programmatic area known as BioPHusion—the identification of critical information requirements (CIRs) and the operationalization of real-time public health information fusion and leadership decision-support activities. Drawing from methodologies for situation awareness used in other domains, we outline the framework being used for the implementation of BioPHusion, including the formalization of information exchange partnerships, systematic information source acquisition, policy development, analysis, research, threat assessments and situational awareness report production. We propose that the framework can be applied to the development of real-time information exchange for situation awareness in other public health practice settings, such as state and local government. And, we suggest that the framework can be used to explore the possibilities around sharing critical information with other components of government involved in the detection of, and response to, public health emergencies.

Keywords: Fusion, public health, bioterrorism, situation awareness.

1 The Need for Real Time Information Exchange in Public Health Practice

A series of public health emergencies over the last 8 years, including the terrorist attacks in the fall of 2001, the spread of Severe Acute Respiratory Syndrome (SARS) and Hurricane Katrina, have highlighted the need for real-time information exchange to enhance government's awareness and understanding of public health events in order for government to prevent or respond to situations as they unfold. However, because the responsibility for public health is shared across levels of government, professional practice and scientific disciplines, the timely sharing of multi-sector, all-hazards, information sharing both is essential and incredibly challenging.

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 1–9, 2008.

Public health situation awareness is needed by public health leaders in three different types of settings that can occur simultaneously or in sequence: 1) pre-event/threat situations where a wide range of public health events and threats are assessed, 2) emergency response situation awareness in which detailed assessments of a specific event or threat and the public health responses to that threat are monitored, and 3) recovery operations during which the on-going mitigation and preventive efforts to a specific event or threat are monitored. Although the techniques for performing information integration, analysis, and reporting of public health-relevant information are very similar across all three, each of these areas has different staffing needs, critical information requirements, and types of output reporting. Pre-event threat assessment includes monitoring a wide variety of information sources to identify potential public health threats or to track existing threats. This activity includes the capture of information, such as media reports, operational reports (i.e. emergency operations center daily reports), or subject matter expert notes, into a tracking database to allow analysts to compile relevant related information over time.

When a public health threat reaches outbreak status requiring emergency response operations, situation awareness operations, which include monitoring not only the spread of disease but also the emergency response preparations, staff and resource deployments, and local public health or other federal agency response activities, are critical to executive decision-making. To support the federal response to public health events, in 2003, CDC formally established the Director's Emergency Operation Center (DEOC) and adapted the Incident Command System (ICS) to meet the unique needs of public health-related emergency responses. The ICS, a detailed management structure first developed for the coordination of federal, state, and local entities in fighting large forest fires and later adapted by the Coast Guard to respond to oil or hazardous cargo spills in the nation's sea ports [1], is now used as a nation-wide standard for emergency response coordination under the National Incident Management System (NIMS) [2]. CDC's adapted version, known as the Incident Management System (IMS), has been incorporated into CDC's disease-specific emergency operations plans to guide the deployment of CDC personnel and to ensure coordinated deployment of other federal support services, such as the distribution of vaccine or anti-viral medications from the Strategic National Stockpile. To address the complexity of the public health response to any event, one of the ways CDC has adapted the ICS for its use was to elevate the Situation Awareness Unit from a sub-component of the Planning Section to a stand-alone section that receives information from across the other components of the ICS and that reports regularly to the federal official in charge of the response. As the event wanes, formal situation awareness activities under the IMS stand down and daily situation awareness is transferred back to CDC programs.

Although the need for public health leaders to possess a shared understanding of ongoing events in order to facilitate decision making and rapid intervention situation awareness is not necessarily new, the need has become more acute with globalization and technological developments. And, the need has been formally recognized through several recent policy developments. The International Health Regulations (IHR), as revised in 2005, represents an international agreement that requires parties to the IHR to develop the surveillance capacity to detect, assess and report to the World Health Organization certain public health events and conditions [3]. The Pandemic and

All-Hazards Preparedness Act (PAHPA), passed in 2006, provides that "the Secretary, in collaboration with State, local, and tribal public health officials, shall establish a near real-time electronic nationwide public health situational awareness capability" [4]. Homeland Security Presidential Directive-21 (HSPD-21), which was issued in 2007, provides for collaboration across HHS and other federal agencies to establish a plan to develop that capability and sets forth other guiding principles for the development of the network, including the need for the network to be flexible, timely, and comprehensive; the need for the network to protect individually identifiable data; and the need for the systems in the network to incorporate data into a nationally shared understanding of current bio-threats and events [5].

2 Public Health Information Fusion: The BioPHusion Model

Recognizing this need for the fusion and sharing of real-time public health information, in 2007 and 2008, the Centers for Disease Control and Prevention (CDC) established the Office of Critical Information Integration and Exchange (OCIIX). The mission of OCIIX is to accumulate and integrate CDC program information and disseminate actionable knowledge on emergent public health events using a meta-analytic approach to ensure all-hazards situation awareness. To accomplish this mission, OCIIX is charged with establishing a new public health fusion center or program at CDC, known as BioPHusion, to "incorporate information from multiple disparate data sources, facilitate the exchange of information across programs, and analyze aggregated interpreted data (information) from existing surveillance systems in order to enhance agency-wide situational awareness both domestically and globally" [6].

Below, we outline the framework for operationalizing this vision through: 1) the development of critical information requirements, 2) systematic information source acquisition, 3) the formalization of information exchange partnerships, 4) analysis, 5) threat assessments, 6) situational awareness report production, and 7) research and policy development.

2.1 Development of Critical Information Requirements

Developing situation awareness capacity for public health begins with identifying the potential diseases, natural disasters, or hazardous exposures that would constitute a public health threat, and determine what the circumstances under which these events should be identified, monitored, and reported and at what level of detail they should be reported. However, because of the spectrum and large number of factors that could affect the public's health, this is not a straightforward or trivial exercise, and these critical information requirements (CIRs) for public health situation awareness must cast a wide net and be flexible, increasing in detail as an event unfolds.

The following high-level CIRs have been established by BioPHusion for its first phase of work to develop a daily situation awareness report for key public health leaders: 1) public health events or threats worldwide; 2) events that indicate a public health event or threat is or may be imminent; 3) threats to CDC staff or resources; and 4) newsworthy public health events, regardless of validity. For the purposes of

BioPHusion's daily situation awareness report, 'public health event' has been broadly defined and includes infectious disease outbreaks, particularly those involving nationally notifiable diseases, potential bioterrorism agents such as anthrax, smallpox, and other "select agents," and accidental exposures such as toxic spills, and natural disasters such as hurricanes and earthquakes. Indications of a potential public health event or one's imminent occurrence if it is not prevented include damage to critical infrastructure or systems, such as water treatment facilities or health care facilities as well as outbreaks that have the potential to overwhelm the healthcare infrastructure.

2.2 Systematic Information Source Acquisition

Ideally, CIRs should guide decisions about the acquisition of information sources that will be routinely monitored to achieve situation awareness. Although public health fusion can be done as BioPHusion currently operates, with primarily open source and publicly available information, such as news media reports, better situation awareness will be achieved when those open sources of information are used along with other sources of public health information, such as summary information from state-level public health surveillance activities and information generated by public health investigations of outbreaks of disease. Public health or biosurveillance surveillance data are collected by states, local health departments, CDC and other federal agencies from a wide range of sources including environmental monitoring systems, animals or vector monitoring systems, individuals, laboratories, medical records, administrative records, police records, and vital records (e.g., birth and death certificates) [7]. Furthermore, these data are collected in a variety of ways (i.e., passive, active, sentinel, special systems, and statistical surveillance) and require flexible approaches to system design and operating procedures [8].

Because data from individual surveillance systems can be incomplete, underreported, or not timely, a combination of sources are needed to have public health situation awareness across the spectrum of diseases and conditions, naturally occurring or intentionally caused, that can impact the public's health [9, 10]. However, there is a real need within public health, especially at CDC, to develop the evidence base to make the case for improvements to those systems so that they facilitate situation awareness. For example, in the spring of 1993 Milwaukee had a large waterborne cryptosporidium outbreak associated with the city's water supply and studies of that event have shown that sales of electrolyte products are good early indicators of respiratory and diarrheal diseases in children and could serve as an earlier signal of an outbreak than hospital diagnoses [11]. Yet, while new systems and sources are attractive, efficiency is also an important developmental consideration in public health and the evidence base for enhancements to existing surveillance systems also needs to be considered [12].

2.3 Formalization of Information Exchange Partnerships

The need for partnerships to achieve broad spectrum subject matter expertise and a continuous flow of information from various sources cannot be overemphasized. The African proverb "if one wishes to travel fast, travel alone but in order to travel far, travel together" is particularly applicable to public health information fusion.

Achieving improved situation awareness at CDC through BioPHusion will only work with support from stakeholders inside CDC and within the broader public health community. Memoranda of agreement or understanding are means by which to establish a clear delineation of working relations between organizations. BioPHusion has memoranda in development with other CDC programs, such as the Director's Emergency Operations Center and the Global Disease Detection Program, as well as other federal agencies that possess key public health-related data, such as the US Department of Agriculture. Informal communications and social networks based on individual relationships can be an important complement to formal lines of communication. Current BioPHusion staff have decades of collective experience at CDC and with states, local governments, non-governmental organizations, and international organizations, such as the World Health Organization. The ideal successful situational awareness network is a hybrid of the informal community 'grass roots' type of information exchange within a formal trans-organizational reporting framework.

2.4 Analysis and Fusion

The types of analysis and fusion conducted in a BioPHusion-type center must reflect the three types of situation awareness needed—pre-event threat situation awareness, emergency or event situation awareness, and event-recovery situation awareness. All three types, but especially pre-event threat analysis and fusion, require that not only must potential health threats be identified and monitored, but analyses and assessments need to be made of the potential threat to at-risk populations, projections made of potential disease spread, and predictions made about the containment or mitigation capacity of public health responses. Real-time fusion is different from, but incorporates, traditional public health meta-analysis. In meta-analysis, combining information is typically thought of with respect to procedures that are intensive over a longer time frame (i.e. weeks, months, or years) than is the case in public health fusion where assessments may be needed in minutes or hours. Meta analysis involves a comprehensive review of the literature and statistical approaches to combine quantitative measures for an aggregate, evidence-weighted conclusion. Public health information fusion for situational awareness involves analogous approaches in a compressed time frame that takes into account both quantitative and qualitative information. It also involves a qualitative research technique known as content analysis, in which the meaning of narratives or groups of narratives is abstracted and themes or meanings are assigned. While reviewing and abstracting all of the many sources of information needed to analyze and identify threats in a compressed timeframe can be challenging, a fully operational situation awareness program should have documentation of logic models, interpretative processes, and analytical techniques in hand to conduct its work.

2.5 Public Health Threat Assessment

Once potential public health events are identified for situation awareness monitoring, there are a variety of additional information sources and analytical services that are necessary to provide a more complete assessment of whether the event is a true health

threat and whether the threat could have severe implications for health if not addressed. Threat assessment begins with documentation of what has been observed, verified or not verified, the importance or irrelevance of information pieces and most importantly, an ongoing dialogue among analysts who may or may not be physically collocated. Wiki or similar technology that draws on the potential of social networks to analyze and understand a problem can be utilized to generate and maintain some of these resource documents [13]. As analysts perform the work, refine analytical skills, obtain new information sources, and develop deeper understanding of nuances of information sources, it is envisioned that they will continuously update this resource knowledge base to accumulate information from which to assess events and threats for their need to be reported.

Specifically, in the BioPHusion model, analysts continuously update the following: 1) situation awareness standard operating procedures or procedures and processes the analysts use to perform daily duties, logic models for decision making, and procedures for activating situation awareness in emergency response; 2) disease characteristics or descriptions of disease processes to support threat assessment, and store information on previous situation awareness assessments of specific threats; 3) information source documentation or description of routine information sources, including known limitations, information lag times, and subject matter expert contacts; and, 4) analytical techniques or processes for performing public health risk assessment, developing advanced disease spread or plume modeling, and creation of advanced GIS information display capacities. These are critical knowledge base areas for the rapid production of threat assessments.

2.6 Situation Awareness Report Production

One of the most critical functions of a situation awareness activity is the report production. Fused and analyzed public health data is of little value if it cannot be accessed and used by public health decision makers to inform a public health response, whether that response is physically placing public health professionals in the field, distributing health messages, teaming with partners in the healthcare or non-profit sectors to deliver services, researching the cause of the event, or developing new policies to minimize the potential of similar future events. Multiple types of information products for different types of public health decision makers may be appropriate. BioPHusion is currently developing public health situation awareness reports for the CDC director, program leadership and selected external partners. In the future, an aim of the BioPHusion program is to develop tools that enable social networking and the creation of on-line communities of public health practitioners to both collect information as well as disseminate situation awareness reports.

2.7 Research and Policy Development

Given the rapidly evolving nature of the information fusion process in public health, each of the components of the BioPHusion model reflected in this paper also have extensive research and policy development needs. The critical information requirements of public health leaders need further exploration. Whether large or small, different public health agencies are usually led by a team, rather than a single

individual. Each member of the team may have different skill sets and perspectives; key informant interviews and evaluation approaches can help to identify the true critical information requirements of the intended audience for public health situation awareness reports. Research into the social networks of the public health community could inform the development of information exchange partnerships as well as approaches to enhancing disease surveillance. There are many statistical, content analysis, information management, and epidemiologic questions that need to be answered in order to do fusion for situation awareness and threat assessment well. Producing effective reports requires further research into how public health decision-makers take in and understand information and use it to make decisions. Public health situation awareness also requires addressing policy barriers to information exchange, exploring the appropriate balance between sharing public health information and protecting the privacy interests and rights of those affected by the information, and facilitating the development of policies and programs that promote that appropriate balance of information sharing and situation awareness.

3 Implications for Other Public Health Practice Settings

BioPHusion seeks to align the unique information repositories within CDC with the public health information gathering that occurs at state and local public health departments by: 1) developing a public health situation awareness report that will be delivered to the CDC director, program leadership and select external partners; 2) developing tools that enable social networking and the creation of on-line communities of public health practitioners; and, 3) expanding programming and application development support to program areas within CDC for information exchange with existing surveillance systems. We suggest that a BioPHusion-type model can be applied to the development of real-time information exchange for situation awareness in other public health practice settings, such as state and local government.

We also propose that this framework can be used domestically by public health practitioners to evaluate the possibility of sharing some or certain critical information with other components of government involved in the response to public health emergencies. The National Strategy for Information Sharing describes the background, current environment, guiding principles and foundational elements for exchanging needed information to detect and prevent terrorism [14]. The Departments of Justice and Homeland Security have developed an extensive national program for supporting the development of state and local law enforcement fusion centers, highlighting the complexity of the landscape and challenges around competing information exchange policies and priorities [15,16]. While data and information sharing to enhance situation awareness and protect the public's health is clearly legitimately needed in circumstances, there also remains much to be worked out regarding how much information is needed, by whom, and for what purposes [17].

BioPHusion, as a program at CDC, is in its infancy and in the initial phases of information source acquisition and report production. The program is likely to evolve dramatically over the next few years and be combined with other agency-wide biosurveillance efforts and approaches to compliance with IHR reporting requirements.

In this paper, we describe the conceptual approach to the components of establishing situation awareness from a federal perspective as a means of inviting dialogue about both the need for public health situation awareness as well as mechanisms for achieving it. Significant cultural, scientific, policy, resource, and communications barriers exist to both achieving shared public health situation awareness at CDC and among the broader public health community. However, anecdotal evidence suggests that the BioPHusion model may be helping to highlight options and opportunities to move public health situation awareness forward.

Author's Disclaimer. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of the U.S. Department of Health and Human Services or the Centers for Disease Control and Prevention.

Acknowledgments. The authors would like to acknowledge the important contributions of the following individuals to the ideas expressed in this paper: Nikki Blye, John Copeland, David Crowder, Vickie Garrett, Janet Heitgerd, Brian Kaplan, Ali Khan, C. Virginia Lee, Richard A. Jones, Arie Managan, Dennis O'Keefe, and Richard A. Scheiber, the Office of the CDC Director, the National Center for Zoonotic, Vector-Borne, and Enteric Diseases, and the many others inside and outside of CDC who have participated in conceptualizing, developing, and supporting the launch of the BioPHusion program.

References

1. Stumpf, J.: Incident Command System: The history and need. Internet J. of Rescue and Disaster Med. 2(1) (2001)
2. National Incident Management System [homepage on the Internet]. : Federal Emergency Management Agency, Washington, DC (cited 2008 September 23), http://www.fema.gov/emergency/nims/index.shtm
3. World Health Organization. International Health Regulations (2005) (cited 2008 September 1), http://www.who.int/csr/ihr/en/
4. The Pandemic and All Hazards Preparedness Act of 2006, Pub. L. No. 109-417
5. Homeland Security Presidential Directive/HSPD-21 [news release]. The White House, Washington, DC (18 October 2007) (cited 2008 August 1), http://www.whitehouse.gov/news/releases/2007/10/20071018-10.html
6. Gerberding, J.: Establishment of the Office of Critical Information, Integration and Exchange (OCIIX) [memoranda]. 23 August 2007. Centers for Disease Control and Prevention, Atlanta, GA (2007)
7. Teutsch, S.M., Churchill, R.E.: Principles and practice of public health surveillance, 2nd edn. Oxford University Press, New York (2000)
8. Centers for Disease Control and Prevention. Guidelines for evaluating surveillance systems. MMWR, 998, 37(S-5), 8-10 (2003)
9. Roush, S., Birkhead, G., Koo, D., Cobb, A., Fleming, D.: Mandatory Reporting of Diseases and Conditions by Health Care Professionals and Laboratories. J. of Am. Med. Assoc. 281(2), 164–170 (1999)
10. Jajosky, R.A., Groseclose, S.L.: Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. BMC Public Health 4, 29 (2004)

11. Hogan, W.R., Tsui, F., Ivanov, O., Gesteland, P.H., Grannis, S., Overhage, J.M., Robinson, J.M., Wagner, M.M.: Detection of Pediatric Respiratory and Diarrheal Outbreaks from Sales of Over-the-counter Electrolyte Products. J. Am. Med. Inform. Assoc. 10(6), 555–562 (2003)
12. National Science and Technology Council. National biological information infrastructure: Protecting against high consequence animal diseases research and development plan for 2008-2012. The White House, Washington, DC (2007) (cited 23 September 2008), www.ostp.gov/nstc
13. Andrus, D., Calvin, T.: The Wiki and the Blog: Toward a Complex Adaptive Intelligence Community. Studies in Intelligence 49(3) (September 2005)
14. The White House. National strategy for information sharing: Successes and challenges in improving terrorism-related information sharing. The White House, Washington, DC (2007)
15. Global Justice Information Sharing Initiative Fusion Center Guidelines [homepage on the Internet]. Department of Homeland Security, Washington, DC (cited 24 September 2008), http://www.it.ojp.gov/topic.jsp?topic_id=209
16. Government Accountability Office. Federal Efforts Are Helping to Alleviate Some Challenges Encountered by State and Local Information Fusion Centers. Report No. GAO-08-35 (October 2007)
17. Goldman, J.: Balancing in a crisis? Bioterrorism, public health and privacy. 38 Journal of Health Law, 481 (2005)

Biosurveillance, Case Reporting, and Decision Support: Public Health Interactions with a Health Information Exchange

Rebecca A. Hills, William B. Lober, and Ian S. Painter

Center for Public Health Informatics, University of Washington

Abstract. This paper describes support for three public health practice domains in demonstrations of a model health information exchange (HIE): biosurveillance, case reporting, and communication from public health to providers through integrated decision support. The model HIE implements interoperability through the use of existing semantic and syntactic standards specified as part of Integration Profiles to support specific data transfer use cases. We implemented these profiles in several public health applications using a service-orientated architecture approach. Methods were validated for each public health domain in national showcase demonstrations. We believe that this work has implications for the integration of public health functions into any HIE, regardless of its architecture, because our informatics methods support a distributed environment. This approach may be extended to strengthen development of the Public Health Grid, a project currently being led by the Centers for Disease Control and Prevention.

Keywords: Health Information Exchange, Surveillance, Case-reporting, Decision Support.

1 Introduction

1.1 Public Health Engagement with the National Health IT Agenda

Both technologies and organizational structures to support Health Information Exchanges (HIEs) have evolved rapidly over the past several years. A variety of other acronyms, including NHII (National Health Information Infrastructure), RHIO (Regional Health Information Organization), SNO (Sub-Network Organization), and RHIN (Regional Health Information Network), refer to variants on the same idea: supporting access to information across organizational boundaries in support of individual and/or population health. HIEs inherently address issues of sharing data across organizations, as well as the semantic and syntactic challenges of this practice, making them potentially invaluable to the increasingly information-rich practice of public health.

In 2004 the United States Department of Health and Human Services (HHS) released The Future of the Public's Health in the 21st Century[1], describing ways in which the United States healthcare system could be rebuilt to take full

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 10–21, 2008.

advantage of health information technology. The report identified four major goals:

1. Inform Clinical Practice: Bringing information tools to the point of care, especially by investing in EHR systems in physician offices and hospitals.
2. Interconnect Clinicians: Building an interoperable health information infrastructure, so that records follow the patient and clinicians have access to critical health care information when treatment decisions are being made.
3. Personalize Care: Using health information technology to give consumers more access and involvement in health decisions.
4. Improve Population Health: Expanding capacity for public health monitoring, quality of care measurement, and bringing research advances more quickly into medical practice.

Goals #1 and #4 clearly identify roles for Public Health. Goal #2 has been broadened to include the goals of security, scalability, and sustainability and is being addressed through the Nationwide Health Information Network (NHIN). Two of NHIN's key activities are promoting the development of HIEs and developing and facilitating the use of standards for clinical exchange [2].

In recent years, HIEs have significantly increased in number and in functionality. Of the seven functional development stages defined by eHealth Initiative, HIEs operating at levels five, six and seven are defined as fully operational [3], demonstrating the transmission of data used by healthcare stakeholders. The 2008 eHealth Initiative annual survey of HIEs [4] identified 42 fully operational HIEs in the United States, up from 32 in 2007 and 26 in 2006. Despite goal #4, to improve population health, only five of the 42 functional HIEs report provision of public health reporting functionality and services.

1.2 Essential Services of Public Health, and the Impact of HIEs

In 1994, the Public Health Functions Steering Committee developed a framework to describe the Essential Services of Public Health [5]:

1. **Monitor** health status to identify community health problems.
2. **Diagnose and investigate** health problems and health hazards in the community.
3. **Inform, educate**, and empower people about health issues.
4. **Mobilize** community partnerships to identify and solve health problems.
5. **Develop** policies and plans that support individual and community health efforts.
6. **Enforce** laws and regulations that protect health and ensure safety.
7. **Link** people to needed personal health services and assure the provision of health care when otherwise unavailable.
8. **Assure** a competent public health and personal health care workforce.
9. **Evaluate** effectiveness, accessibility, and quality of personal and population-based health services.
10. **Research** for new insights and innovative solutions to health problems.

Many of the public health practices and processes used to provide these services may be substantially enhanced by changes in the availability and flow of information, both at the population and individual levels. Three examples of potentially improvable practices that support one or more of these Essential Services are: biosurveillance, case reporting, and the communication of alerts and guidelines from public health to clinical providers through integrated clinical decision support. Population surveillance and case reporting are important components of the first two Essential Services: monitoring health status of a community, and diagnosing and investigating health problems within a community. Population surveillance also plays a vital role in #10, research. Immunization decision support and communication provide benefits falling into Essential Services #6 and #7, enforcement and linking people to health services.

Examining these three practices in the context of an HIE highlights an interesting spectrum of requirements. First, biosurveillance, reporting, and decision support require, respectively: unidirectional flow of de-identified data, unidirectional flow of identified data, and bidirectional flow of clinical data and tailored recommendations. Second, there is a wide range in technological maturity of these practices in an HIE context. Third, biosurveillance is mostly conducted within public health agencies, while reporting and integrated decision support require case-level interaction with care providers. Fourth, while all three practices are tied to essential services, they are viewed differently within the public health community. Notifiable condition case reporting is a core process in health departments, while the return on investment of biosurveillance remains controversial, but both are fairly well understood. In contrast, decision support driven by public health is a novel approach to distributing public health alerts and guidelines, with very few real-world implementations to demonstrate its concepts and value.

1.3 Demonstrating Public Health Interaction with an HIE

The Integrating the Healthcare Enterprise [6] (IHE) initiative was undertaken in 1998 by healthcare professionals and industry to improve interoperability of healthcare information systems. In pursuit of this goal, IHE promotes the coordinated adoption and use of existing healthcare IT standards through an ongoing collaborative process involving multiple players. In 2005 the Healthcare Information Technology Standards Panel [7] (HITSP) was created as part of an effort by HHS to promote interoperability in healthcare by harmonizing health information technology standards. HITSP does not create standards, but instead identifies existing standards for particular use cases. IHE has supported this effort, using the IHE framework to demonstrate HIE capabilities across clinical and population health use cases. The standards identified by HITSP have been incorporated into Interoperability Profiles developed by IHE.

The phases of IHE's coordination and adoption process are defined as: problem identification, integration profile specification, implementation and testing, and integration statements and requests for proposals. During the implementation and testing phase, vendors employ the integration profiles and participate

in face-to-face testing activities with other vendors. After the "Connectathon" testing activities, IHE sponsors Interoperability Showcases at both the Health Information and Management Systems Society [8] (HIMSS) and Public Health Information Network [9] (PHIN) annual conferences. The purposes of these showcases are to bring awareness of IHE activities to both the HIMSS and PHIN communities and to provide a live demonstration of a model HIE.

IHE's model HIE, as demonstrated during the showcases, can serve as a proxy for performance at the level of a fully operational organization (levels 5, 6, and 7, as defined by the eHealth Initiative). Participation by a diverse set of vendors, as well as the focus on scenarios, standards, and integration profiles making use of standards, make the model HIE well-suited for testing and demonstration of public health's role in information exchange.

During the past three years, our research group has participated in five IHE Showcases where health information exchange capabilities were demonstrated at national health IT (HIMSS) [10] and public health informatics (PHIN) [11] meetings [12]. Our role was to highlight increasingly rich information interactions between the HIE and public health. We organized our demonstrations first around biosurveillance, and then biosurveillance and case reporting together and most recently we added communication from public health to clinicians through decision support services. We successfully integrated both existing and new tools into the HIE framework developed as part of the IHE initiative.

2 Methods

The primary goal of the IHE Showcases is to illustrate technical use cases for information exchange through scenarios involving three to six vendor systems. A story line is created to give a realistic feel to the scenario and to engage conference attendees in specific capabilities of the model HIE. Groups of attendees are given an orientation to IHE and then taken on walking tours through the scenario, observing the interchange of information in different vendor systems. We have typically played the role of public health, showing HIE information in systems built to illustrate specific practices within public health organizations. We will describe the IHE framework, specifics of the three scenarios we helped to develop, and the framework we used to develop the purpose-built public health applications.

2.1 The IHE Technical Framework

The IHE Technical Framework is essentially an integration guide, detailing standards-based transactions between information systems [13]. IHE's Integration Profiles (IPs) build on the Technical Framework, defining specific actors and identifying transactions to realize use cases that address specific needs. For example, one IP used in our demonstration is Patient Identifier Cross-Reference (**PIX**), which implements a master patient index. The Cross-Enterprise Document Sharing (**XDS**) profile makes use of ebXML [14], SOAP [15] and HL7 [16]

Clinical Document Architecture (**CDA**) specifications to define clinical document sharing. The Retrieval Forms for Data Capture (**RFD**) profile uses XForms [17] technology to enable the gathering of form data from within an application, for submission to an external location.

2.2 Biosurveillance Scenario

The scenario defined for the surveillance demonstration describes a patient with severe flu-like symptoms who visits his primary care provider. The provider sends samples to the lab for influenza type and subtype testing, and refers the patient to the emergency department. As depicted in Figure 1, the visit summary, the referral, and the lab orders and results are available as documents through the HIE, using the XDS profile. In the emergency department the attending physician reviews the referral document from the primary care provider and the laboratory results. Public health is able to see both aggregate surveillance results and individual patient results for A(H5N1) positive samples.

The HIE established a Shared Document Resource and granted access to authorized providers as well as to the public health actor. Also, as represented in Figure 1, public health conducted regular polling of the Shared Document Resource, looking for laboratory documents of a specific type, e.g., positive labs for A(H5N1). To accommodate the population perspective of public health a new type of query was established that allowed public health to look for documents in a shared document resource or repository without specifying a patient identifier. This allowed public health to fulfill monitoring, surveillance and event management roles by retrieving population level data, while preserving security and auditing frameworks as no identifiers were transferred to public health, and logs of each "public health query" transaction were recorded. In recent demonstrations we created a public health dashboard to demonstrate visualizations of the surveillance data. The dashboard provided access to tools for data management, classification, analysis and characterization as well as visualization and results (Figure 1). These data representations were made possible through integration of an existing web services framework (Shoki) [18]. The XDS system, its scalability, and integration with an IBM Research epidemiological modeling tool have been previously described [19].

2.3 Case Reporting Scenario

In the case-reporting scenario, also represented in Figure 1, a 28-year-old male patient returns to his provider for the results of an HIV test. The provider reports the positive HIV results to public health by accessing and auto-populating an HIV Case-Report Form through the Electronic Medical Record (EMR) and then submitting the form to Public Health. These steps were implemented using the RFD profile. A vendor system acted as a Forms Manager, one of the actors in the RFD profile, hosting the reporting forms and making the URLs available to vendors wishing to add case-reporting functionality to their systems. After retrieving a case-report form, based on physician action, a vendor EMR system

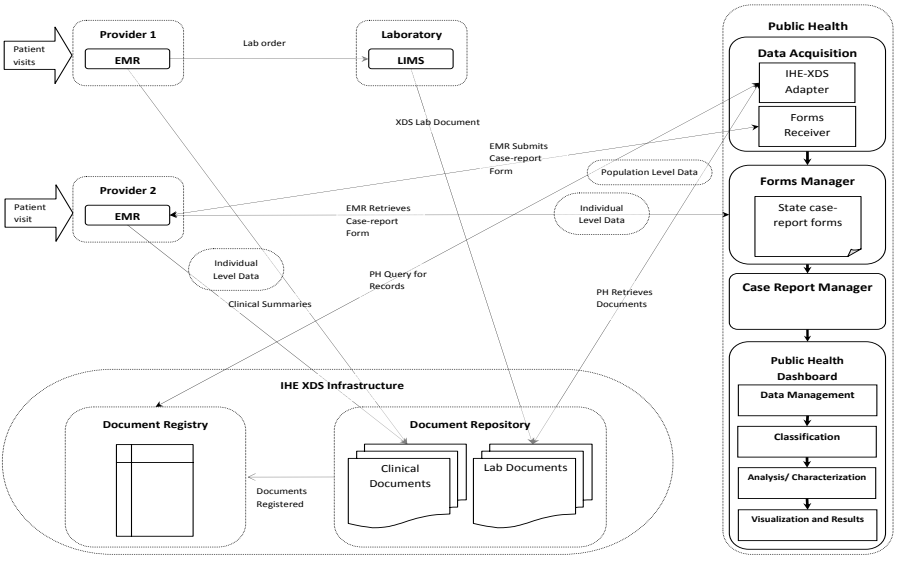


Fig. 1. Prototype Health Information Exchange Showing Public Health Interactions for Biosurveillance and Case-Reporting

used data from the patient record to populate fields on the form. Optionally, supplemental information was completed by the provider before the form was submitted to public health. The submission of the case report is represented in Figure 1 by the arrow between Provider 2 and the Public Health Forms Receiver. Public health received the reporting form, which was then stored in a local case-report repository. Through a link to the public health dashboard, new HIV case-reports were displayed as alerts. A public health case-report manager accessed the list of case-reports, and was able to view details of the case and annotate the form to reflect the ongoing case investigation.

Case-report forms for HIV were developed using the official case-report forms for the states of Massachusetts and Washington. Xforms technology was used to create forms similar in layout to the paper forms for the two states. The Shoki web services framework [18] was used to make available tools for data management, classification, analysis and characterization and visualization and results.

2.4 Immunization Scenario

The immunization scenario shown in Figure 2 describes visits by a young child and his parents to a clinician in the child's state of birth, and then later to a clinician in another state. It then addresses the need to ensure that, with respect to the current requirements of that particular state, the child's immunizations are up to date when he begins school. At the first visit, the provider registers the patient with the HIE and sends a record of the administered immunizations

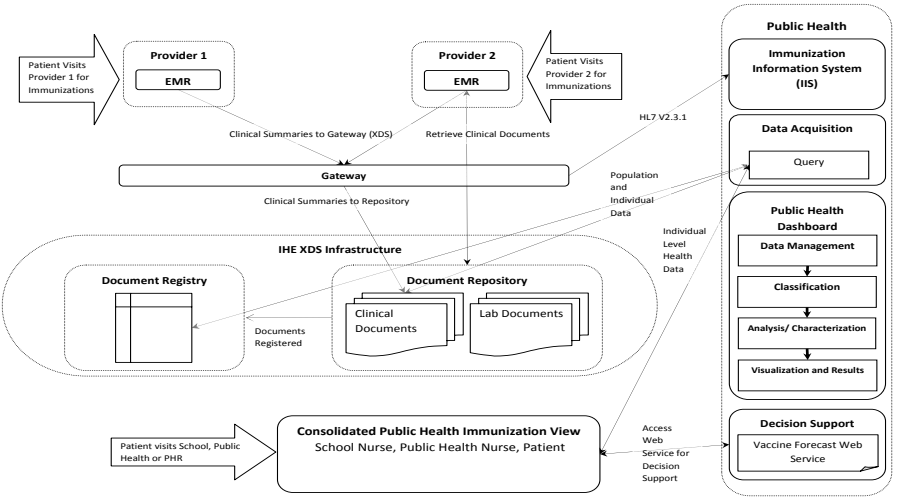


Fig. 2. Prototype Health Information Exchange Showing Public Health Interactions for Immunization

as a CDA document to the HIE's document repository via a gateway. This gateway allows the provider to update both the state Immunization Information System (using an HL7 V2.3.1 message) and the document repository (an XDS document) with the transmission of a single message. The child's parents come to an appointment with a provider in a different state several years later, but are unable to provide an immunization record. This provider is able to query the HIE to find previous records for the child. These records show that the child is in fact due for several more immunizations. These immunizations are administered and the record of this visit is also sent to the HIE.

The final part of this scenario demonstrates two roles for public health as the child is enrolled in kindergarten. Again, the parents are not in possession of the child's immunization record but a nurse, with access to the HIE (via the consolidated immunization view represented in Figure 2), is able to locate the child's immunization records and display an immunization history. This limited access would be appropriate for either a limited scope clinical provider, such as a school nurse, a public health immunization clinic, or the parent themselves, with appropriate authorization through an HIE-compliant personal health record. The second role of public health direct, patient-specific, communication with the provider or patient is illustrated by the provision of a decision support web service for vaccine forecasting. This web service performs de-duplication and validity tagging of immunizations submitted and provides an immunization schedule based on the current immunization recommendations.

In order for this seamless communication between providers, public health and web services providers to take place, the existence of a standard method of representing immunization information is essential. IHE's Immunization Registry

Content Profile [20], although still in draft form, was used successfully during the demonstrations to enable the sharing of immunization information. This profile defines standard messaging, document and web services formats for immunization data exchange among Immunization Information Systems and EMR systems, HIEs and public health.

2.5 Integration of an Existing Web Services Framework

The existing web services framework, which we used to build public health example applications for the Biosurveillance and Case-Reporting scenarios, is the Shoki framework [18]. Shoki consists of services grouped into four areas, representing the four core functions of surveillance systems: data management; classification; characterization and analysis; and visualization and reporting. Within these four focus areas, Shoki offers pluggable components allowing public health to make use of diverse data sources, in this case, using data from the HIE. During each of the showcase demonstrations we participated in, Shoki was easily integrated into the IHE framework by writing wrappers to map IHE representations of data to Shoki's services. This has allowed flexible acquisition of incoming data from the commercial systems participating in IHE.

3 Results

Our participation in the HIMSS 2006, 2007 [12] and 2008 [21] and PHIN 2007 and 2008 showcases brought the voice of public health into the scenario development process, served to inform both the clinical and the public health communities about IHE activities, demonstrated important public health practices to conference attendees, and helped the IHE community to understand that certain use cases and requirements of public health differ from those of clinical users.

The IHE showcases schedule regular tours for conference attendees. Each tour explores one of the scenarios step-by-step, making stops at each of the participant stations, e.g., EMR system vendors, lab system vendors, infrastructure providers and public health. The size of the tour groups reflected the relative sizes of the conference (HIMSS is a much larger conference than PHIN) and interests of the attendees reflected the different demographics of the groups.

Approximately 15 half-hour tours were given during the $3\frac{1}{2}$ days of the showcase at the HIMSS conferences. Most conference attendees at HIMSS are associated with the clinical community and the showcase brought awareness to that community of the role and needs of public health within an HIE. At the two PHIN conferences, where $2\frac{1}{2}$ days of demonstrations took place, approximately 10 tours were given and a large number of individuals visited outside of the scheduled tour times. Questions and discussion initiated by visitors at PHIN indicated a strong interest in IHE activities and the model HIE, as well as interest in the possibility of extending the IHE model to support bidirectional communication between public health and providers, and specifically in demonstrating this communication in the next Showcase.

We demonstrated effective surveillance on both clinical and laboratory documents by periodically polling document registries to identify and retrieve relevant information. The structure of these queries is unique because they lack patient identifiers. Clinical users were required to provide patient identifiers to retrieve documents and no IHE mechanism existed prior to our participation to conduct a query looking for a set of documents matching some clinical criteria, e.g., Influenza A. To accommodate the needs of biosurveillance, a new public-health specific query was developed and implemented for use in the biosurveillance scenario.

Submission of case-reports using the RFD profiles successfully demonstrated a timely and efficient method for clinicians to comply with public health regulations. With the cooperation of EMR vendors and the standardization of a method for retrieving forms, we demonstrated a streamlined method of case-reporting that made both high-quality data available to public health rapidly and that introduced little interruption to clinical workflow. The integration of the Shoki web-services framework demonstrated a range of analytic and visualization techniques made possible by the use of a single technical framework.

The public health and the XDS communities were successfully linked and a model for bidirectional communication was demonstrated by the use of the Immunization Registry Content Profile. Retrieval of immunization data from repositories of clinical documents was straightforward. The use of the retrieved data for personal health records, public health use or school use was demonstrated. Making use of the document and patient registries, and patient demographic queries allowed the document repositories to compliment immunization registries. Bidirectional communication was implemented through the novel invocation of an external, decision support web-service [22] based on current Advisory Committee for Immunization Practice recommendations.

4 Discussion

We were able to demonstrate, alongside commercial vendors of EMR, laboratory, and other health applications, the potential benefits to public health of the standards harmonization efforts led by IHE. These efforts have made possible improved communications between public health and both providers and laboratories, which in turn make possible integrated surveillance systems that increase the ability of public health to identify and address threats in a timely manner. The potential for public health uses of clinical data continues to be better understood, and although this demonstration used only one of the several frameworks for an HIE, the ideas presented generalize to other HIE frameworks, as well as other methods of connecting public health to clinical providers, including Grid technologies.

Providing patient-tailored decision support is an example of public health-to-provider communication (other examples are case identification, information requests during investigations, and distribution of other alerts and guidelines). It is rare, outside of community clinics or other situations where a public health agency is the direct provider of clinical services, for public health to have permission to directly insert guidelines into a medical record. Even where this capacity

may exist, the technical obstacles are daunting. The IHE framework allows patient-tailored decision support and guidance to be made available to clinicians regardless of which EMR they use. This capacity offers perhaps the greatest potential of all three use cases we have demonstrated.

With the help of public health informatics practitioners and researchers, the use of both an information exchange framework and the service oriented architecture (SOA) methodology has the potential to make secondary use of clinical data for public health a consistent reality. The model HIE used during the showcases is one of many possible frameworks for the connections between providers, laboratories and public health entities; however embracing an SOA methodology is likely to increase the possibility of reuse of tools developed regardless of which model is used for information exchange. SOA is not limited to HIEs, and the pluggable components demonstrated in our work could be incorporated into a Grid infrastructure in order to expose data and services for use by other organizations on the Grid. There is great potential for expanding the Shoki tools demonstrated at the IHE Showcase. For instance, one might expose or retrieve data via an adapter, which would bind certain XDS transactions to OGSA-DAI [23] Grid queries.

The approach we are taking to Grid integration has led us to a broader definition of Shoki service categories, to reflect a spectrum of services beyond those initially described for surveillance [18]. The original categories of data management: classification, characterization and analysis, and visualization and reporting, have been expanded to include the use cases of decision support/communication and case reporting. We have reorganized the services used for these IHE Showcases into: Distributed Query/Synthesis, Data Integration/Management, Data Quality, Case Detection, Classification, Characterization/Analysis, Visualization/Reporting, Security/Management, and Messaging/Alerting. These categories correspond approximately to those of the National Public Health Toolkit recently proposed by CDC's National Center for Public Health Informatics.

Clearly HIE's can be of benefit to public health, but public health's unique requirements are not always recognized. Public health can use interoperability standards to monitor population health trends, improve both the efficiency and rate of reporting of notifiable conditions, and to provide input to clinical care. Public health routinely collects data at various levels of aggregation- the level of the individual for case reporting, and at higher levels of aggregation for population surveillance. At the lowest level of aggregation, i.e., notifiable conditions, the reporting of clinical data is usually legally mandated, and patient identification is essential for public health follow-up. At higher levels, such as population surveillance, data collection is legally mandated in some jurisdictions, but often no law requires the collection of data. In this context, patient identifiers are not essential to public health; in fact, transferring identifying information may be detrimental to the community's trust of public health institutions. Concerns with privacy and trust initially led the IHE framework to support only queries based upon specific patient identifiers. The higher aggregation levels required for public health to provide certain services indicate a need for data exchange

that does not exist in the clinical context for which most HIE's were developed. Thus, it is not adequate for public health to simply wait for and then use HIEs. One of the most important lessons learned as a result of participation in these demonstrations has been that partnerships between public health, private institutions and vendors are needed to make certain that public health needs are considered as the HIE infrastructure becomes a reality.

5 Conclusions

We demonstrated to both a public health and a clinical audience, that integration of public health functions within a model HIE is possible, desirable, and achievable. In the future, expanding the demonstrations to include example practices impacting more of the ten essential services will bring attention to other needs and constraints on public health and clinical providers and the way they interact. It is our belief that most data-centric functions of public health will benefit from the increased efficiency, improved data quality, and increased availability of data made possible by HIE integration. Leveraging the data resources made available through HIEs can help identify and address quality of care issues, can benefit population health research, and can serve to more fully integrate disease registries. We see these potential benefits as significant and hope to continue engaging IHE, the HIE community, and public health, and to move on to real-world implementation of the concepts presented here.

Only with the implementation of the IHE framework in a real HIE environment, rather than the model of the IHE Showcase, will we be able to discover and accurately describe the technology and policy challenges to adoption of the frameworks being developed by IHE. While the IHE profiles are adopted by new vendors each year, it is also important to examine legacy systems and practices in clinical and public health settings. The methods necessary to retrofit these systems to enable participation in an HIE based on this framework will be important to ensure maximum participation in HIEs in the future.

Public health has increased its use of information technology to acquire, analyze and report on data of interest to public health practitioners. In this time of rapid development and change in information systems it is essential that public health be a part of the decision-making process in order to efficiently and effectively access and use the data being made available by laboratory and EMR systems as they are integrated into HIE's.

Acknowledgements. We would like to gratefully acknowledge support from the National Institutes of Health/National Library of Medicine/Robert Wood Johnson Foundation (Grant # 3 T15 007442-06S1) and the Centers for Disease Control and Prevention through their support of the University of Washington Center for Excellence in Public Health Informatics. Thanks also to all of the IHE Showcase participants especially Alean Kirnak at Software Partners and Lori Reed-Fourquet.

References

1. Institute of Medicine (IOM). The Future of the Publics Health in the 21st Century. Committee for Assuring the Health of the Public in the 21st Century. National Academies Press, Washington DC (2002)
2. Office of the National Coordinator for Health Information Technology. Summary of Nationwide Health Information Network (NHIN) Request for Information (RFI) Responses (2005) (September 1, 2008), http://www.hhs.gov/healthit/rfisummaryreport.pdf
3. eHealth Initiative State of the Field: Whats Happening?, http://www.ehealthinitiative.org/2007HIESurvey/stateOfTheField.mspx
4. Marchibroda, J.: Bringing Health Information Exchanges and Public Health Together. In: Bringing health information exchanges and public health together forum at the Public Health Information Network Annual Conference (2008)
5. Public Health Functions Steering Committee. The Public Health Workforce: An Agenda for the 21st Century. Full Report of the Public Health Functions Project, U.S. Department of Health and Human Services (1994), http://www.cdc.gov/od/ocphp/nphpsp/essentialphservices.htm
6. Integrating the Healthcare Enterprise (IHE), http://www.ihe.net/
7. Healthcare Information Technology Standards Panel , http://www.hitsp.org/
8. HIMSS, http://www.himss.org/
9. Public Health Information Network, http://www.cdc.gov/PHIN/
10. HIMSS Conference, http://www.himssconference.org/
11. PHIN Conference, http://www.cdc.gov/phinconference/
12. Rodriguez, C.V., Lober, W.B., Sibley, J., Webster, E., Painter, I., Karras, B.T.: Integrating public health applications with commercial EMRs. AMIA Annual Symposium Proceedings (2007)
13. Integrating the Healthcare Enterprise Technical Framework, http://www.ihe.net/About/ihefaq.cfm
14. ebXML (Electronic Business Using Extensible Markup Language), http://www.ebxml.org/
15. Simple Object Access Protocol (SOAP), http://www.w3.org/TR/soap/
16. Health Level Seven (HL7), http://www.hl7.org/
17. W3C Forms Working Group, http://www.w3.org/MarkUp/Forms/
18. Lober, W.B., Drozd, D., Lumley, T., Sebestyen, K., Painter, I.: An open source web services toolkit for event detection algorithms. Adv. Dis. Surv. 1, 78 (2006)
19. Carmeli, B., Eschel, T., Ford, D., Greenshpan, O., Kaufman, J., Knoop, S., Ram, R., Renly, S.: Public health affinity domain: a standards-based surveillance system solution. In: Zeng, D., Gotham, I., Komatsu, K., Lynch, C., Thurmond, M., Madigan, D., Lober, B., Kvach, J., Chen, H. (eds.) BioSurveillance 2007. LNCS, vol. 4506, pp. 147–158. Springer, Heidelberg (2007)
20. IHE Immunization Registry Content Profile, http://wiki.ihe.net/index.php?title=ImmunizationRegistryContent
21. Painter, I.S., Hills, R.A., Lober, W.B., Randels, K.M., Sibley, J., Webster, E.: Extending Functionality of and Demonstrating Integrated Surveillance for Public Health within a Prototype Regional Health Information Exchange. In: AMIA Annual Symposium (2008)
22. Software Partners Vaccine Forecasting Module, http://www.matchmerge.com/en/vaccine-forecast-decision-support-service.html
23. OGSA-DAI, http://www.ogsadai.org.uk/

Bio-surveillance Event Models, Open Source Intelligence, and the Semantic Web

Nancy Grady[1], Lowell Vizenor[2,3], Jeanne Sappington Marin[4], and Laura Peitersen[5]

[1] Science Applications International Corporation, 301 Laboratory Road, Oak Ridge, TN 37831
[2] Computer Task Group, 701 Ellicott Street, Suite B2-146, Buffalo NY 14203
[3] Formerly of Ontology Works, 3600 O'Donnell St. Suite 600, Baltimore, MD 21224
[4] Science Applications International Corporation, 1001 Research Park Boulevard Suite 302, Charlottesville, VA 22911
[5] Science Applications International Corporation, 1235 South Clark Street, Suite 610, Arlington, Virginia 22202
{gradyn,marinjs,peitersenl}@saic.com, lowell.vizenor@ctg.com

Abstract. Surveillance applications to monitor health-related data have matured rapidly over the last several years. A newly emerging development is an emphasis on harvesting and evaluating the timely but potentially inaccurate information present in unstructured sources such as Internet news feeds and sites. An important development for the surveillance on both structured and unstructured datasets is the exchange not of the primary datasets that feed these systems, but of the evaluated results of such analysis. This paper introduces recent work addressing a model for the recording and tracking of events and for the dissemination of information about these events to other agencies. It will introduce a structured relational database model for events, an ontology for infectious disease events, and a semantic web representation. The strengths and weaknesses of the three approaches and future directions will be discussed.

Keywords: Biosurveillance, open source intelligence**,** event model, ontology, semantic web.

1 Introduction

Traditional automated surveillance systems rely on the integration and analysis of streams of structured data from public or private sources and the development of algorithms to detect anomalous activity within those streams. These systems operate upon the structured primary or secondary source data that is transmitted to the analytical application. An emerging emphasis is on the development of data fusion or open source intelligence centers, which seek to harvest and integrate information from the most timely but most unreliable of sources, the open Internet. Some systems, such as Clark Freifeld and John Brownstein's HealthMap [1], provide for automated harvesting of events; others, through automated systems, assist in the manual creation of events ([2]-[4]).

The outcome of surveillance across both structured and unstructured datasets by applications and by analysts is the identification of events of significance. While there are standards for the representations and transmission of primary and secondary

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 22–31, 2008.

laboratory and clinical data, such as the Health Level Seven (HL7) [http://www.hl7.org/], message protocol and the American Health Information Community (AHIC) Minimum dataset [http://www.hhs.gov/healthit/standards/resources/], there has not been an emphasis on the automated exchange between systems of the vetted results of surveillance. As more and more local and regional surveillance systems mature, it is important that the vetted results of surveillance across systems can be exchanged and be actionable for other organizations.

An actionable specification of the outcome of biosurveillance would need to minimally include a description of the dataset processed, the activity and affected region, the significance of the anomaly, and an evaluation of the confidence in the accuracy of the information and in the analytics techniques used to identify the anomaly. Biosurveillance systems should produce as an end product a tertiary record of events of significance, whether natural or man-made, that can be analyzed, tracked, and exchanged with other surveillance systems and with stakeholders to increase the coverage and effectiveness of surveillance across the nation.

The National Bio-surveillance Integration System (NBIS) ([5], [6]) is an application designed to facilitate the identification of events of national significance through the harvesting of open source information and to facilitate the collaboration on and dissemination of the details of these events with a broader community.

2 Related Work

Work has progressed on ontology-centered or knowledge-based bioterrorism surveillance, through an ontological framework for describing and matching data and methods in a surveillance application ([7], [8]). A simple ontology of events has been used for events found through natural language processing on texts ([4], [9]). An event calculus for tracking epidemic spread has been developed [10] in terms of the descriptions of spatio-temporal objects. A number of ontologies have been developed and are available at the Open BioMedical Ontologies site [http://www.obofoundry.org/].

To stimulate discussion in the field on a semantic exchange of the structured details about events found as the outcomes of biosurveillance, this paper will describe the implementation within NBIS of an event model for recording and tracking significant events. An "event" can be an elemental occurrent or a compound set of related occurrents. The next section will describe three representations of an event model. The implementation in the NBIS system and in a semantic web demonstration will then be presented, followed by a discussion of the strengths and weaknesses of the different approaches and finally some future directions.

3 Methodology

There are three main approaches to recording and tracking structured details about events; to create a set of tables in a relational database that have fields to record the relevant information; to use a semantic relationship model through the representation of the data, using an ontology as a schema; to represent the details of an event as a Semantic Web markup within unstructured texts.

3.1 Relational Event Model

An event model was created for the NBIS 2.0 application using a relational database for population of the details of events of significance, through both manual and automated data entry. While the initial emphasis on the event model was for infectious diseases, the need for tracking events in other domains was considered in the design. The main categories that were created for tracking an event consisted of an event header; the host species; the agent, disease, vector and reservoir involved; the location; the source of the information; and the evidence or findings.

Header. Each event has an overall description of the event and its stage in the analytical process. The first database table describes the unique event identifier, a title, a user label for convenience in grouping, and a version number for tracking the history of the data entry.

In addition to these identifiers, there are overall descriptions of the event. The descriptions are the domain (human, animal, plant or the environment); the agent category (chemical, biological, radiological or climactic); the relevant threat scenario (for example, pandemic influenza, foot and mouth disease or food contamination); the overall starting and ending dates; and the background and significance of the event to place it in context.

Finally, there are workflow indicators to track events through the analytical and publication process.

Table 1. Event workflow elements

Element	Description
Validity	Is the event valid, or was it generated to test the system or generated during a simulation or table-top exercise
Status	Is the event record in the initial stages of description; has it been accepted as a significant event; has it been modified; or has it been rejected as a significant event
Surety	The confidence in the event as suspected, confirmed, or denied
Criticality	Understanding the significance of the impact of the event, with indications such as low, medium, or high
Change	In what way the event has changed since its last report, such as new information having been provided that changes the original understanding of the event; it has been updated to reflect new developments; it has been judged to have a greater impact; or it has been judged to be lessening in impact
Publication	Whether the event is currently under distribution to stakeholders as an active event

The model consists of one main table to hold this descriptive information. All other categories of details are held in separate tables to allow a one-to-many relationship, so multiple hosts or locations, for example, could be used to describe an event.

Hosts. The second category for an event description contains the details of the host of the event. If the host involves an organism, the host can be described using a common name, and also the official name from an organism taxonomy, such as the NCBI [11]. If the host is human, then gender can be specified, along with upper and lower bounds for ages to handle aggregate data. Additional information about hosts are the multiple symptoms and/or risk factors that are present.

Agents. The category of the agent for the event is given in the header. Additional details are included in this separate table to give the specific organism, chemical, nucleotide or climactic agent. Allowance was made to be able to specify a toxin, for example, as the agent separate from the host organism generating it.

Diseases. Often in open sources, a distinction is not made between the agent causing a disease and the disease itself. As there is not a one-to-one relationship between them, a separate container was created to hold the common name for the disease as well as an element for the standard name.

Vectors and Reservoirs. To enable a fuller description of the context of an event, the common and standard names for the vector and reservoir can be recorded. This provides context when an agent that is perhaps endemic in one region is present in a new region due to the transmission by a different vector, such as apparently occurred in the 2005-2006 Reunion Island Chikungunya outbreak.

Locations. The location for an event can consist of the named location or a latitude and longitude; a description of the extent of the event (such as a county, state, country, or continent); and any relevant geographic features, such as landform, population density or development.

Sources. In surveillance on structured datasets, the source of the data is typically fixed and well understood. In surveillance using open source information, the origins of the details of a breaking event must be described. Not only must the publisher be cited, but also the identification and role of the authority being quoted in the publication need to be recorded. There is furthermore a need for analysts to specify their confidence in the reported results. In a recent open source intelligence study [12], a "chain of denial" corpus of national reports in several countries showed a not uncommon pattern of official denials in the face of an eventually confirmed outbreak.

Findings. The most difficult aspect to model is the description of the findings or observational evidence of the event. Most importantly a finding for an outbreak or case consists of the specific lab test and result, the date of the sample or of the test, and the laboratory performing the test. An additional element allows for comments concerning the specificity and accuracy of the test. Since many of the fields are similar, the finding table can also be used for other types of observational evidence related to the event. It is difficult to have sufficient fields for recording information without having analysts have to resort to free text entry fields, making accurate aggregation, query and retrieval more difficult.

The tables for the header, hosts, agents, diseases, locations, vectors and reservoirs, locations, sources, and findings provide the fields to enter and track the details about

an event. These details allow querying, aggregation and summarization over multiple events, or the dissemination of specific details about a single event.

3.2 Ontology for Infectious Diseases

All surveillance systems provide for the storage and integration of data, but typically only deal with a reasonably small number of data types and domains. The integration among datasets requires configuration tables or views that specify the key columns that can be used to join across tables and the columns that represent the same information. As the number of datasets and domains they represent increase, the maintenance of such mapping tables becomes unwieldy. For the purpose of integrating large, heterogeneous datasets, a more robust approach is the semantic integration of datasets through the use of an ontology. The ontology provides a single, unifying framework wherein it is possible to (1) define in a very precise manner the entities and relations involved in a given domain(s), (2) encode subject matter expertise in the form of rules, (3) link datasets with other knowledge sources such as clinical terminologies and biomedical ontologies, and (4) provide a global query schema for data retrieval.

An ontology for infectious diseases was prototyped within the NBIS 2.0 application to integrate datasets from differing domains. It was initially targeted at published avian influenza case details and was expanded to provide for the storage and integration of the existing NBIS data feeds from across domains, providing the objects, events, attributes and relations needed to minimally describe the structured data feeds. The infectious disease ontology was developed within the framework of Basic Formal Ontology (http://www.ifomis.org/bfo), an upper-level ontology that underwrites a principle-based approach to ontology design. In addition, a number of reasoning modules were added to the ontology to support temporal [13] and spatial [14] reasoning. Finally, a modified version of the ISO Standard Common Logic (ISO/IEC 24707:2007) was used to encode the ontology content. (Note: Automated translation methods have been developed to translate the infectious disease ontology into OWL [http://www.w3.org/TR/owl-ref/] a Semantic Web-based ontology language.)

The NBIS ontology consists of a class hierarchy, where the two highest-level classes are *Continuant* and *Occurrent*. These are technical terms that mark a fundamental way that entities exist in time. Continuants are entities that preserve their identity over time, even as they gain and lose parts. The NBIS ontology further divides the Continuant class into two disjoint subclasses: *Object* and *Object Attribute*. Objects are the bearers of qualities such as diseases and symptoms and the entities that participate in events (e.g. the organisms that are involved in an outbreak). Examples of Continuants in NBIS include Organisms, Organizations (e.g. Hospitals and Government Agencies), Geographic Regions (e.g. Countries and Cities), and Artifacts (e.g. Documents).

Object Attributes are entities that depend for their existence on the objects that bear them. Examples of object attributes include states, roles, qualities, functions, etc., and can be organized into taxonomies. Examples of object attributes in NBIS include object qualities such as diseases, symptoms, gender and age as well as roles such as pathogen and host.

Since we are using the general term "event" to mean the anomalous activity discovered through biosurveillance, we have replaced the commonly used ontology term “event” with the technical term "occurrent." Occurrents are dynamic entities that have temporal parts. Examples of occurrents include the transmission of infectious agents, the spreading of a disease and the onset of symptoms. Table 2 represents a sample of the NBIS ontology hierarchy.

Table 2. NBIS Ontology (Sample)

- Continuant
  - Object
    - Organism
    - Organization
    - GeographicalRegion
  - Object Attribute
    - State
      - BeingInfected
    - Role
      - Pathogen
      - Host
    - Quality
      - Disease
      - Symptom
- Occurrent
  - ClinicalEvent
    - DiagnosticProcedure
    - LaboratoryProcedure
    - Hospitalization
  - PublicHealthEvent
    - Infection
    - Onset
    - Recovery
    - Death

Besides the subclass relation that forms the backbone of the NBIS ontology, there are also a number of associative (i.e. non-hierarchical) relations that link entities together. Each relation in the NBIS ontology is given a signature that identifies the types of entities that are related to one another through a given relation. Here we classify a number of relations used in the NBIS ontology in terms of the high-level classes: Object, Object Attribute, and Occurrent. For example, the familial relation *parentOf* would be classified as an Object-to-Object relation since it relates two objects.

Table 3. NBIS Relations (Examples)

- Object-to-Object
  - locatedIn
  - spatialPartOf
  - memberOf
- Object-to-ObjectAttribute
  - hasDisease
  - hasSymptom
  - hasRole
- Occurrent-to-Object
  - hasHost
  - hasPathogen
- Occurrent-to-Occurrent
  - temporalPartOf
  - precedes

The NBIS classes and relations provide the basis for the creation of a number of rules that make it possible to make connections between heterogeneous data feeds and semantically integrate those datasets.

3.3 Semantic Web Representations of Events

The vision of the Semantic Web [15] is an extension to the current Web to tag information to make it accessible to automated processing. Currently, biosurveillance information is obtained through text mining techniques that harvest content from semi-structured Web pages, as for example, in the HealthMap system [1], or it is obtained directly through structured data services such as RSS (en.wikipedia.org/wiki/RSS) feeds. Many sites export RSS data, for example, from EpiSpider [http://www.epispider.org/], GDACS [http://www.gdacs.org], or RSOE [http://visz.rsoe.hu/alertmap/index.php?lang=].

There are a number of semantic frameworks being developed by different communities under the Linking Open Data umbrella [16]. These frameworks describe the contents of Web pages in much the same way as an ontology, but use the elements of objects and properties. Data is described through the use of RDF (www.w3.org/TR/rdf-primer) triples, which specify a relationship: Subject -> predicate -> object, which can be used to describe the full ontology in OWL.

The NBIS relations can be easily mapped into this triplet framework where subject and object are taken from the ontology's objects, attributes, and occurrents. The predicate is taken from the set of relations, e.g. Host -> hasDisease -> Disease.

An advantage of this approach is that RDF schemas can be published and easily incorporated into other systems. Tools exist for the creation and automatic processing of such datasets. In addition the properties can be tagged within unstructured texts, for example, in a semantic wiki, that allow structured data queries over the texts.

4 Implementation

The event model was implemented in the NBIS 2.0 application with a relational database representation. Open source articles were harvested from the Internet from a number of sites and news feed queries for streaming to system users. Conceptual categorization was used to indicate if an article referred to current events or better fit other categories, such as historical retrospectives, corporate product announcements, or research grants. The article list was presented through the user interface created using a wiki. Users can scan article titles, sources, and categories, and select articles that describe an event of interest. Entity identification was used to extract location, agent and disease mentions from selected articles to pre-populate an event record. Additional wiki extensions provided for viewing or editing the elements in an event record. A workflow extension was created to allow analysts to route the events for additional research, rejection, or approval to publish. The final wiki extensions allowed analysts to map specific event elements to extract into templates for creating wiki pages with the information desired for publication.

An ontological schema for infectious diseases was prototyped within the NBIS project to transform structured data feeds for storage in a relationship database. Common logic queries can retrieve event records according to the semantic meanings of the elements, independent of the data feeds the data came from, and through the contextual data relationships, for example, to allow ease of queries such as "…from any country adjacent to this country."

An additional demonstration case study has been developed to put a corpus of articles on recent Chikungunya outbreaks into a semantic wiki. The semantic wiki provides for summarization of the objects in the articles, with easy querying for other pages containing the same objects or predicates.

5 Discussion

There are strengths and weaknesses to the differing approaches to event representation. A relational database model imposes a rigid structure on the model, making changes difficult. The inflexibility of the model potentially encourages users to want to use free-form text in fields, rather than limited code sets, reducing the utility of queries over events. The one-to-many table structure means that separate events need to be created if, for example, there are different vectors at different locations. Relational structures are easier to implement and are more familiar to analysts, but are difficult to change in production systems.

An ontological representation must be created by experts to ensure there are no circular definitions. The ontology makes it much easier to incorporate contextual information for richer and yet simpler queries, since relationships can be traversed to generate query results. The query language and relationship framework is, however, not as familiar to users and requires significant user training.

The semantic web representation has the advantage of a simpler representation than an ontology, flexibility in adding relationships, and easier reuse of other semantic representations. The tools for the manipulation of the schemas are more commonly available. It has an advantage of allowing annotation directly within the unstructured text, useful for open source intelligence systems. As the tools and their usage matures, this approach will allow structured queries over unstructured material, due to the semantic tagging, and data interchange among systems through Web services. It is more difficult, however, to construct aggregate data from a semantic representation.

6 Conclusions and Future Work

This work has used three different approaches to generate specific models of events within biosurveillance that allow for the editing, versioning, workflow and publication of these events. These models have been used to describe events harvested from open source intelligence and to integrate structured data from different domains. The relational model is deployed in the production NBIS system. The ontological data integration and semantic web representation are being evaluated for utility of use.

A significant issue in event representation is the continuing need for standard vocabularies and contextual datasets. While NCBI has an extensive taxonomy of organisms, it is, of course, incomplete. Good contextual datasets are difficult to obtain for the wide range of data needed for comprehensive biosurveillance.

Open source data is the most timely, but also the most unreliable of information. Additional work is needed to standardize the representation of the provenance [http://twiki.ipaw.info/bin/view/Challenge/] of the data and the assessment of its reliability for better evaluation of the events.

This preliminary work indicates the potential of semantic data representations for the conduct of biosurveillance, providing integration of disparate datasets and providing structured data through annotation of unstructured articles. The principal challenge to ensuring the scalability of a system lies in ensuring the completeness of the semantic representations, identifying important contextual datasets, and encoding the surety and provenance of the analysis.

The most important next step is in community discussion of an event model for the specification of standards for the interchange of information between biosurveillance systems. The advantages of having the experts closest to the data provide the analysis to specify the events are clear, as is the need to exchange information between surveillance systems covering different regions, datasets, and domains.

A future technical direction is to integrate the ontology and semantic web representations with other ontological development, an example being an infectious disease ontology at the Open Biomedical Ontologies (http://www.obofoundry.org/).

A more significant challenge for the future will be the construction of reasoning engines to process these exchanged events, being described [17] as "Higher Order Mining." The data mining community has dealt with the need to combine the results of multiple models generated by the same algorithm on the same data through voting schemes (known as bagging or boosting). Less well developed is the need to combine the results of different algorithms applied to the same data (known as stacked generalization). A reasoning system over events will be difficult, since these events are, in effect, the output scores of different models over different datasets, with the output reduced to a score of one for the event region and zero elsewhere. Reasoning over events will be reduced primarily to overlap detection and change detection as the event extent and severity change over time. The ability to reason over the correlation between events will be a challenge if only the vetted event details are exchanged.

As Chute stated in a recent editorial [18] "…the emerging dependency of the health sciences on increasingly practical semantic technologies to organize and leverage these vast information resources is now unquestioned." In this paper, we argue that these technologies also need to be applied to event modeling to facilitate information exchange among biosurveillance systems and practitioners.

Acknowledgments. This work was supported by the Department of Homeland Security, Chief Medical Officer's Office of Health Affairs, National Biosurveillance Integration System 2.0, under contract HSHQDC-06-D-00026, and by the support of Science Applications International Corporation. The authors acknowledge extensive discussions on the relational event model with Lynne Hendricks, Teresa Quitugua, Michelle Podgornik, and Tom Slezak, and the guidance of the National Biosurveillance Integration Center Director, Eric Myers.

References

1. Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. J. Am. Med. Inform. Assoc. 15, 150–157 (2008)
2. Wilson, J.M., Polyak, M.G., Blake, J.W., Collmann, J.: A Heuristic Indication and Warning Staging Model for Detection and Assessment of Biological Events. J. Am. Med. Inform. Assoc. 15, 158–171 (2008)

3. Centers for Disease Control and Prevention, Office of Critical Information Integration and Exchange (CVHG), http://edocket.access.gpo.gov/2008/pdf/E8-10986.pdf
4. Grady, N.W., Jones, J.H., Vizenor, L., Dudley, J.P., Marin, J.S., Havey, A., Sabo, T., Champoux, R., Hogan, R.L., Beal, D., Peitersen, L., Orr, J., Kropp, K., Smith, K.L.: Data Integration and Analytics within the National Bio-Surveillance Integration System. Advances in Disease Surveillance 4, 94 (2007)
5. Smith, K.L.: Statement to the House, Subcommittee on Prevention of Nuclear and Biological Attack, Creating a Nation-wide, Integrated Biosurveillance Network, Hearing (May 11, 2006), http://chs.clientapp2.com/hearings/viewhearing.aspx?id=30
6. Smith, K.L.: Keynote Presentation at the NSF Biosurveillance workshop. New Brunswick, NJ (2007)
7. Curbezy, M., O'Conner, M., Pincus, A., Musen, M.A., Buckeridge, D.L.: Ontology-Centered Syndromic Surveillance for BioTerrorism. IEEE Intel. Systems 20, 26–35 (2005)
8. Buckeridge, D.L., Graham, J., O'Connor, M.J., Choy, M.K., Tu, S.W., Musen, M.A.: Knowledge-Based Bioterroism Surveillance. In: Proc. AMIA Symp., pp. 76–80 (2002)
9. Kawazoe, A., Chanlekha, H., Shigematsu, M., Collier, N.: Structuring an event ontology for disease outbreak detection. BMC Bioinformatics 9, S8 (2008)
10. Chaudet, H.: Extending the event calculus for tracking epidemic spread. Artificial Intelligence in Medicine 38, 137–156 (2006)
11. National Center for Biotechnology Information Taxonomy, http://www.ncbi.nlm.nih.gov/Taxonomy
12. Dudley, J.P., Marin, J.S., Peitersen, L.: OSINT Using Text Reports of Official Comments Official Actions & Anomalies during HPAI Outbreaks. Technical Report for NBIS
13. Allen, J.F.: Maintaining knowledge about temporal intervals. Communications of the ACM, 832–843 (26/11/1983)
14. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. In: Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning, pp. 165–176. Morgan Kaufmann, San Mateo (1992)
15. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American Magazine (May 17, 2001)
16. W3C Sementic Web Education and Outreach Interest Group Community Project, http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
17. Roddick, J.F., Spilopoulou, M., Lister, D., Ceglar, A.: Higher Order Mining. In: SIGKDD Explorations, vol. 10, pp. 5–17 (2008)
18. Chute, C.G.: Biosurveillance, Classification, and Semantic Health Technologies. J. Am. Med. Inform. Assoc. 15, 172–173 (2008)

Foresight China II: Identification and Detection of Infectious Diseases

Jiayuan Feng[1], Jianshi (Jesse) Huang[1], and Angus Nicoll[2]

[1] Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China
[2] European Centre for Disease Prevention and Control, Health Protection Agency, London School of Hygiene and Tropical Medicine, UK
pumcjesse@yahoo.com.cn

Abstract. With the growing threat of emerging and re-emerging infectious diseases, which spread more easily and quickly worldwide resulted from the globalization, it is necessary to be capable to forecast significant changes of infectious diseases for emergency preparation purpose. Foresight China Project aims at exploring a new method by identifying the drivers of infectious diseases and predicting the trends of these drivers.

Keywords: Infectious diseases, surveillance, foresight, drivers.

1 Background

Today's world is a small village. Globalization is by no means a new phenomenon, transcontinental trade and the movement of people began at least 2000 years ago since the ancient Silk Road trade route. The global spread of infectious diseases followed a parallel course, to some extent, it's the epitome of globalization. Now, after two millennia, human pathogens are experiencing another bonanza from a new era of globalization, which is characterized by faster travel over greater distance and global trade. Under this circumstance, what happens in China could have significant impact on other places or countries and vice versa.

With the growing threat of emerging and re-emerging infectious diseases, it is necessary to forecast significant changes of infectious diseases for emergency preparation purpose. However, "It is very difficult to make predictions, especially about the future". The global scientific community is faced with the huge task of assessing the possible future impacts of global change drivers on infectious diseases.

The existing methods used to predict future trends in infectious diseases are mostly quantitative predictions, including predictions and modeling. However, the quantitative methods, which are based on historical data and use mathematics and statistics to make estimations, are mainly fit for short-term prediction within 5 years. Out of this scope, a mass of variables and conditions can no more be estimated, in which case the confidence interval will expand immeasurably [1]. Besides, the quantitative methods

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 32–41, 2008.

are limited to specialized pathogens, e.g. HIV/AIDS and Measles, they do not apply well to unknown or emerging pathogens [2].

Without some moderately accurate predictions or at least early warning, we cannot have a safe global village. To develop a holistic approach to risk assessment of infectious diseases under global change, we need to change our view from the direct risk assessment of specific infectious disease outbreaks, to focus on predicting the trends of risks of infectious diseases occurrence, and then trends of diseases. Foresight China is a novel and simple approach to forecasting future trends in drivers and families of infectious diseases, needs for surveillance and public health preparedness.

2 Method

Based on literature review and experts consultation, Foresight method identified a series of families of risks for the study of possible trends in those drivers.

2.1 Foresight China I

Foresight China I was under performance from 2004 to 2006, supported by Foresight Funding, enacted by PUMC and HPA. The Basic Risk Model for Infectious Disease Risks was adapted from UK colleagues, and 36 leading Chinese experts were consulted. Some important factors affecting future risks were identified, including governance and social cohesion, demography and population change, conflict, Technology & Innovation and their governance, agriculture and change of land use, economic factors, trade and market related factors, transport and tourism, human activity and social pressure[3-5].

2.2 Foresight China II

From the result of Foresight China I, both HPA and China consider this program useful and meaningful, so the Foresight Funding determined to do a follwoing research, that's the Foresight China II. In July 2007, we carried out the Foresight China II with improved methodology, supported by the British Embassy Beijing and enacted by PUMC.

From lessons learnt from Foresight China I, we made some improvement in our methodology. First, we did systematic literature review, based on which we identified the scientific evidence for 12 families of drivers of infectious diseases occurrence; Second, we analyzed the existing infectious disease surveillance systems in 4 countries (China, Britain, America and Japan) to identify if they capture data on these drivers; Third, we defined each driver in the questionnaire which would achieve improved consensus. At last, we expanded our consultation to 181 leading Chinese experts to confirm the improvement opportunities to assess feasibility of the opportunities and to predict trends of the drivers in China.

3 Result

3.1 Systematic Literature Review

As people's knowledge of disease turned from biomedical mode into the "biology-medicine-society-environment" medical mode, role of "environment" factor in the "epidemiological triangle", which affects the epidemical occurrence and prevalence, arouses people's attention. When environment changes, pathogens and hosts also change. The research studies the major influencing factors of epidemical occurrence and prevalence from these three aspects, which are "pathogen, host and environment".

Based on the literature review, despite of the nine family drivers identified in Foresight I, we identified another three family drivers on infectious diseases[6-8] , which are environment related factors, iatrogenic related factors, animals and plants related factors. Finally, there are 12 families and 50 elements identified as drivers of infectious diseases occurrence.

3.2 Extensive Analysis of Existing Infectious Disease Surveillance Systems in 4 Countries

3.2.1 The Identify of 'Key Surveillance Infectious Diseases'

Our research compared the infectious disease surveillance systems in China, Britain, America and Japan, including the contents under surveillance and the data they have collected. The infectious disease surveillance systems of Britain, USA and Japan are chosen to compare with that of China, because Britain and USA are western developed countries with relative integrated surveillance systems, and Japan is an Asian country with similar culture and geographical environment to China. Their achievements in the infectious disease surveillance system are worthy for us to study.

Because of the presence of an enormous number of infectious diseases, it is neither necessary nor practicable to monitor all these diseases in most countries or areas, especially in those with quite limited resources. Surveillance system of 'key surveillance infectious diseases', that is, infectious diseases with major public health significance in certain country or area, should be established primarily.

According to the six principles[9] introduced by WHO on the definition of key surveillance infectious diseases and 37 notifiable infectious diseases in China, we identified 18 key surveillance infectious diseases, which are Tuberculosis \ Measles\ Typhoid and paratyphoid\ Malaria\ Newborn tetanus\ Rabies\ Gonorrhea\ Syphilis\ HIV/AIDS\ Hemorrhagic fever with Renal Syndrome\ Encephalitis B\ Epidemic cerebrospinal meningitis\ Leptospirosis \ Plague\ Dengue fever\ Bacillary and amoebic dysentery\ Viral Hepatitis and Bird flu.

3.2.2 Analysis of Infectious Disease Surveillance Systems in China, Britain, USA and Japan

Theoretically, as mentioned in many references [10-12], the key information that should be collected by infectious disease surveillance system are: demographic data, disease morbidity or mortality, investigation data of influencing factors, record of intervention measures, topic-based investigation report.

But in fact, from the analysis of surveillance status of 18 key surveillance infectious diseases in four countries, we find that the present infectious disease surveillance mostly stalled at the biomedical model, which emphasizes the passive surveillance of patients and pathogens, or, collection of demography and morbidity/mortality. However, investigation data of influencing factor is seldom collected.

3.3 Leading Experts Consultations in 12 Areas

3.3.1 Characteristics of the Experts

Table 1. Foresight China II: Identification and Detection of Infectious Diseases: characteristics of the experts consulted

		Number	Proportion (%)
Age(year)			
	<40	23	12.7
	40-60	137	75.7
	>60	21	11.6
Area			
	North China	108	59.7
	East China	32	17.7
	Middle China	7	3.9
	Southwest	8	4.4
	Northwest	10	5.5
	Northeast	7	3.9
	South China	9	5.0
Title			
	Advanced	116	64.1
	Associate-advanced	59	32.6
	Middle	6	3.3
Education Lever			
	Doctor	52	28.7
	Master	51	28.2
	Bachelor	72	39.8
Major			
	Management	6	3.3
	Public Health	97	53.6
	Clinical Medicine	15	8.3
	Basic Medicine	12	6.6
	Agriculture	15	8.3
	Economic and Trade	8	4.4
	Tourism and Traffic	5	2.8
	Environment	8	4.4
	Other	15	8.3
Years on this major			
	Below 10	8	4.4
	10-19	68	37.6
	20-29	79	43.6
	30 and above	25	13.8

3.3.2 Results

Area 1 Governance and social cohesion

Table 2. Foresight China II: Identification and Detection of Infectious Diseases "Governance and social cohesion" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Bio-security governance	12(85.7)	2(14.3)	14(100)	0	14(100)	0
Social cohesion	7(50.0)	7(50.0)	6(42.9)	8(57.1)	4(28.6)	10(71.4)
Illegal practices	8(57.1)	6(42.9)	8(57.1)	5(35.7)	7(50.0)	6(42.9)
international/national/regional interactions affecting governance	14(100)	0	13(92.9)	0	12(85.7)	1(7.1)
lack of interaction between policy and regulatory agencies	14(100)	0	10(71.4)	3(21.4)	8(57.1)	5(35.7)
marginalization of some groups specify	12(85.7)	2(14.3)	13(92.9)	1(7.1)	13(92.9)	1(7.1)
political leadership	13(92.9)	1(7.1)	7(50.0)	6(42.9)	4(28.6)	10(71.4)

Area 2 Demography and population change

Table 3. Foresight China II: Identification and Detection of Infectious Diseases "Demography and population change" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Immigration/migration urbanization	13(86.7)	2(13.3)	14(93.3)	0	13(86.7)	2(13.3)
Aging population	9(60.0)	6(40.0)	7(46.7)	8(53.3)	9(60.0)	4(26.7)
gender imbalance	9(60.0)	6(40.0)	9(60.0)	6(40.0)	8(53.3)	6(40.0)
occupation changes	11(73.3)	4(26.7)	13(86.7)	2(13.3)	12(80.0)	1(6.7)
education	15(100)	0	11(73.3)	4(26.7)	10(66.7)	1(6.7)

Area 3 Conflict

Table 4. Foresight China II: Identification and Detection of Infectious Diseases "Conflict" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Difficulties in maintaining administrative systems	15(93.8)	1(6.3)	14(87.5)	2(12.5)	13(81.3)	3(18.8)
Movement of refugees	16(100)	0	11(68.8)	5(31.3)	10(62.5)	6(37.5)

Area 4 Technology &Innovation and their governance

Table 5. Foresight China II: Identification and Detection of Infectious Diseases "Technology &Innovation and their governance" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Ability to control infections control strategies	15(93.8)	1(6.3)	16(100)	0	11(68.8)	4(25.0)
New techniques of Information transmit	16(100)	0	16(100)	0	13(81.3)	3(18.8)
Drug or Pesticide Resistance	14(87.5)	2(12.5)	16(100)	0	13(81.3)	3(18.8)
Faster identification and diagnostics of organisms	13(81.3)	2(12.5)	11(68.8)	3(18.8)	11(68.8)	4(15.0)
Use of new medicine methods and technologies	8(50.0)	8(50.0)	12(75.0)	4(25.0)	15(93.8)	1(6.3)
Survival of patients with chronic disease	8(50.0)	8(50.0)	9(56.3)	7(43.8)	11(68.8)	5(31.3)

Area 5 Agriculture and land use change

Table 6. Foresight China II: Identification and Detection of Infectious Diseases "Agriculture and land use change" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
landform and physiognomy	12(100)	0	10(83.3)	2(16.7)	12(100)	0
dense model of agriculture	9(75.0)	3(25.0)	9(75.0)	3(25.0)	9(75.0)	3(25.0)
Changing Patterns of Land Use	11(91.7)	1(8.3)	10(83.3)	2(16.7)	10(83.3)	2(16.7)

Area 6 Economic factors

Table 7. Foresight China II: Identification and Detection of Infectious Diseases "Economic factors" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Economics development level	14(100)	0	11(78.6)	3(21.4)	7(50.0)	7(50.0)
Income disparity	13(92.9)	1(7.1)	9(64.3)	5(35.7)	7(50.0)	7(50.0)
Poverty and Malnutrition	13(92.9)	1(7.1)	11(78.6)	3(21.4)	11(78.6)	2(14.3)
Unemployment	10(71.4)	4(28.6)	6(42.9)	8(57.1)	8(57.1)	6(42.9)

Area 7 Trade and Market related factors

Table 8. Foresight China II: Identification and Detection of Infectious Diseases "Trade and Market related factors" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Behavior, Structure of Markets and changing pattern of Trade	15(100)	0	15(100)	0	14(93.3)	1(6.7)
Demands for exotic products	13(86.7)	2(13.3)	15(100)	0	15(100)	0
Illegal trade	14(93.3)	1(6.7)	14(93.3)	1(6.7)	8(53.3)	6(40.0)
Trade Barriers	9(60.0)	6(40.0)	12(80.0)	3(20.0)	10(66.7)	4(26.7)

Area 8 Transport and Tourism

Table 9. Foresight China II: Identification and Detection of Infectious Diseases "Transport and Tourism" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Movement of People, animals, micro-organisms	18(100)	0	18(100)	0	18(100)	0
Tourism	16(88.9)	2(11.1)	17(94.4)	1(5.6)	17(94.4)	1(5.6)

Area 9 Human activity and social pressure

Table 10. Foresight China II: Identification and Detection of Infectious Diseases "Human activity and social pressure" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Changes in Sexual Practices	14(93.3)	1(6.7)	12(80.0)	3(20.0)	11(73.3)	4(26.7)
Changing Lifestyle	13(86.7)	2(13.3)	14(93.3)	1(6.7)	11(73.3)	4(26.7)
Public perceptions	11(73.3)	3(20.0)	11(73.3)	4(26.7)	13(86.7)	2(13.3)
Demands for more Healthy Food	13(86.7)	1(6.7)	12(80.0)	3(20.0)	11(73.3)	4(26.7)
Media Reporting	10(66.7)	4(26.7)	5(33.3)	10(66.7)	5(33.3)	10(66.7)
Faith	12(80.0)	3(20.0)	14(93.3)	1(6.7)	14(93.3)	1(6.7)

Area 10 Environment related factors

Table 11. Foresight China II: Identification and Detection of Infectious Diseases "Environment related factors" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Environment pollution	15(100)	0	12(80.0)	3(20.0)	10(66.7)	5(33.3)
Safe of water and foods	13(86.7)	2(13.3)	14(93.3)	1(6.7)	13(86.7)	2(13.3)
Climate warm up	12(80.0)	3(20.0)	11(73.3)	4(26.7)	9(60.0)	4(26.7)
Nature disaster	15(100)	0	15(100)	0	13(86.7)	0

Area 11 Iatrogenic related factors

Table 12. Foresight China II: Identification and Detection of Infectious Diseases "Iatrogenic related factors" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Overcrowding in hospital	14(87.5)	2(12.5)	14(87.5)	2(12.5)	13(81.3)	3(18.8)
Hospital acquired infection	14(87.5)	2(12.5)	14(87.5)	2(12.5)	14(87.5)	2(12.5)

Area 12 Animals and plants related factors

Table 13. Foresight China II: Identification and Detection of Infectious Diseases "Animals and plants related factors" expert opinion

elements	Driver of infectious diseases?		Need to surveillance?		Feasibility to surveillance?	
	Yes	No	Yes	No	Yes	No
Animal movement	15(100)	0	15(100)	0	14(93.3)	1(6.7)
Change of animal species	15(100)	0	14(93.3)	1(6.7)	13(86.7)	2(13.3)
Change of animal husbandry methods	13(86.7)	2(13.3)	14(93.3)	1(6.7)	12(80.0)	2(13.3)
genetically modified crops or animals	10(66.7)	5(33.3)	10(66.7)	5(33.3)	11(73.3)	4(26.7)
gene polymorphism decrease	15(100)	0	11(73.3)	4(26.7)	10(66.7)	5(33.3)

Table 14. Foresight China II: Identification and Detection of Infectious Diseases: feasibility to surveillance of 12 family drivers

12 family drivers of infectious diseases occurrence	Doable	Can be done	Consider latter
1.Governance and social cohesion	Bio-security governance	Illegal practices	Social cohesion
	international/national/regional interactions affecting governance	lack of interaction between policy and regulatory agencies	political leadership
	marginalization of some groups specify		
2.Demography and population change	Immigration、migration and urbanization	Aging population	
	occupation changes	gender imbalance	
		Education	
3.Conflict	Difficulties in maintaining administrative systems	Movement of refugees	
4.Technology &Innovation and their governance	New techniques of Information transmit	Ability to control infections control strategies	
	Drug or Pesticide Resistance	Faster identification and diagnostics of organisms	
	Use of new medicine methods and technologies	Survival of patients with chronic disease	
5.Agriculture and land use change	landform and physiognomy		
	dense model of agriculture		
	Changing Patterns of Land Use		
6.Economic factors	Poverty and Malnutrition	Economics development level	
		Income disparity	
		Unemployment	
7.Trade and Market related factors	Behavior, structure of markets and changing pattern of Trade	Illegal trade	
	Demands for exotic products	Trade Barriers	
8.Transport and Tourism	Movement of People, animals, micro-organisms		
	Tourism		
9.Human activity and social pressure	Public perceptions	Changes in Sexual Practices	Demands for more Healthy Food
	Faith	Changing Lifestyle	
		Population immune status	
10.Environment related factors	Safe of water and foods	Environment pollution	
	Nature disaster	Climate warm up	
11.Iatrogenic related factors	Overcrowding in hospital		
	Hospital acquired infection		
12.Animals and plants related factors	Animal movement	genetically modified crops or animals	
	Change of animal species	gene polymorphism decrease	
	Change of animal husbandry methods		

4 Discussion

In information collection, in order to ensure quality of the research, as well as acquire useful information, researchers choose experts strictly according to the standard. As the research does not refer to sensitive personal questions, and the research is one of the important work that the nation matters, experts showed their interests in the research and been supportive and cooperative, which make sure that the result is authentic and reliable. To avoid information bias caused by the interviewer's misunderstanding, there are extra explanations for some special content in the consultative questionnaire for experts, and the interviewers are trained, examined and given a uniform notice.

At present, people are facing new situation of epidemical occurrence and prevalence, especially new epidemics are found continually. Human behavior, society and natural environment differ from the past [13-14]. Present epidemic monitoring system apply biomedical mode, mainly depends on case report and laboratory diagnosis, can not discover epidemical occurrence and prevalence in time. As people's acquaintance of disease turned from biomedical mode into the "biology-medicine-society-environment"

medical mode, the view of influencing factors of epidemical occurrence and prevalence becomes to be the affection to epidemical occurrence and prevalence caused by economy and social development.

Pathogen, host and environment in the "epidemiological triangle" are the biological basis of epidemic occurs and prevail in public. America started the early warning of the West Nile Virus infection according to death monitoring of the animal host and mediator, explained that epidemic monitoring system help monitor influencing factors of epidemical occurrence and prevalence, find abnormal conditions, send out early warning and start epidemic control.

Therefore, this research suggests China put feasible factors among the major influencing factors of epidemical occurrence and prevalence into epidemic monitoring system. Except reports from hospitals and laboratories, related department should put more attention into the change of the major influencing factors of epidemical occurrence and prevalence, as well as analyze it, so as to find and control epidemic occurrence and development in time, enhance the early warning function of the epidemic monitoring system.

References

1. Xie, Z.H., Huang, J.S.: Discussion on predicting infectious diseases. Chinese General Medical Journal 11(1), 85–87 (2008)
2. Garnett, G.P.: The transmission and control of Ebola -virus. In: Ludwig Mathematical models for the spread of infectious diseases, pp. 104–121
3. Mary, E.W.: Travel and the Emergence of Infectious Diseases. Emerging Infectious Diseases 1(2), 39–46 (1995)
4. Stephen, S.M.: Factors in the Emergence of Infectious Diseases. Emerging Infectious Diseases 1(1), 7–15 (1995)
5. Forrest, D.M.: Control of imported communicable diseases: preparation and response. Can. J. Public Health 87(6), 368–372 (1996)
6. Robert, W.S.: Global Change and Human Vulnerability to Vector-Borne Diseases. Clinical Microbiology Reviews 17(1), 137–174 (2004)
7. Joseph, N.S., Manish, A.D., Levy, K., et al.: Environmental Determinants of Infectious Disease: A Framework for Tracking Causal Links and Guiding Public Health Research. Environmental Health Perspectives 115(8), 1216–1223 (2007)
8. John, A.C., David, R.M., Michael, G.B.: Emerging Infectious Diseases in an Island Ecosystem. The New Zealand Perspective. Emerging Infectious Diseases 7(5), 767–772 (2001)
9. WHO. Communicable disease surveillance and responsible system. Guide to monitoring and evaluating (2006)
10. Halperin, W., Baker Jr., E.L.: Public Health Surveillance. Van Nostrand Reinhold, New York (1992)
11. Langmuir, A.D.: The surveillance of communicable diseases of national importance. NAJM 268, 182–191 (1963)
12. Qu, S.: Disease surveillance. Epidemiology, 4th edn. People's Medical Publishing House, Beijing (1999) (in Chinese)
13. Cohen, M.L.: Changing patterns of infectious disease. Nature 406, 762–767 (2000)
14. Morens, D.M., Folkers, G.K., Fauci, A.S.: The challenge of emerging and reemerging infectious diseases. Nature 430, 242–249 (2004)

Public Health Preparedness Informatics Infrastructure. A Case Study in Integrated Surveillance and Response: 2004–2005 National Influenza Vaccine Shortage

Ivan J. Gotham[1,2], Linh H. Le[1], Debra L. Sottolano[1], and Kathryn J. Schmit[1]

[1] New York State Department of Health, Empire State Plaza, Albany, New York 12236 USA
[2] Department of Biometry and Epidemiology, School of Public Health, University at Albany, State University of New York 12222 USA
{ijg01,lhl02,dls20,kjs05}@health.state.ny.us

Abstract. Effective Public Health Emergency Preparedness (PHEP) requires integrated information systems supporting key PHEP activities, including surveillance, alerting, situational awareness, emergency planning and response, resource assessment and management. These systems are optimized when embedded within an informatics framework supporting a community of information trading partners engaged in routine health information exchange. Seasonal influenza (flu) in the USA typically peaks in January-February and accounts for over 200,000 hospitalizations and 36,000 deaths annually. Vaccination is the primary method of prevention and the optimal preseason time for vaccination is September-November. The October 5, 2004, announcement of significant influenza vaccine shortfalls triggered a national PHEP event, requiring a full array of integrated and heightened PHEP activities at the state and local levels. The presence of an established integrated informatics framework for health information exchange in NY State conveyed significant advantages in advanced preparedness and just-in-time response to the event. This paper describes how the framework supported and enhanced the efficacy of NY's response to a real-life hazard, details related performance metrics, and presents lessons learned from the response.

Keywords: Health Preparedness Informatics Surveillance Response Influenza vaccine shortage.

1 Introduction

Public Health Emergency Preparedness (PHEP) is a process of reaching a sustainable state of "readiness to act" as part of the essential public health activities practiced by health departments daily [1,2,3]. Effective PHEP requires integrated information systems supporting a spectrum of key *routine* public health activities, including surveillance, event detection, alerting, situational awareness, emergency planning and response, resource assessment and management. These systems are optimized when embedded within an established informatics framework supporting a broad-based community of health information trading partners engaged in routine information exchange [1,2,3].

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 42–55, 2008.

An established integrated informatics framework for health information exchange conveys significant advantages in advanced preparedness and just-in-time response to actual health events. This work describes how such a framework supported and enhanced the efficacy of a state's response to a real-life and real-time PHEP event: the 2004 national influenza vaccine shortage. We also present details on related performance metrics and lessons learned from the event response.

1.1 Background and Events Leading to the National Influenza Vaccine Shortage PHEP Event

Seasonal influenza (flu) in the USA typically accounts for over 200,000 hospitalizations and 36,000 deaths annually. Vaccination is the primary method of prevention and the optimal time for vaccination is September-November, in advance of the typical peak in flu activity in January-February [4]. The 2003-2004 flu season, one year prior to the shortage, was atypically severe. The onset of peak influenza activity occurred over November-December 2003. Influenza-related morbidity indicators (e.g., hospitalizations) and mortality were 2-3 times that observed in 3 previous flu seasons. The season was also associated with a 2-3 fold increase in influenza-related pediatric hospitalizations and deaths [5,6,7]. The publicity surrounding the deaths and the severity of the flu season increased demand for vaccine, resulting in some localized shortages [8]. In October 2004, influenza-associated pediatric deaths became a nationally notifiable condition [6]. The details leading up to the national vaccine shortage are available in federal reports [9]. In brief, it began on October 5, 2004, when production problems in a major vaccine manufacturer, Chiron, cut the U.S. supply of vaccine in half. The heightened demand for vaccine following the 2003-2004 season and the timing of the shortage led to the proverbial 'perfect storm' of emergency conditions.

The event response covered the spectrum of PHEP activities. The CDC's plan consisted of two phases [9]. Phase I began on October 12, 2004, when the CDC released limited vaccine orders, previously placed with alternate manufacturers, for providers and health care facilities according to estimates of risk group needs. In November 2004, CDC Phase II required States to place statewide orders for vaccine to meet priority risk group needs unmet by Phase I and other deliveries made prior to the shortage. States had to activate emergency response plans to: 1) assess vaccine availability through previous orders and CDC Phase I; 2) assess unmet priority risk group vaccine needs across health care facilities, updating as the situation changed; 3) analyze and estimate vaccine to be ordered through CDC Phase II, updating as the situation evolved; 4) develop a statewide allocation and distribution plan for LHDs and health care facilities, based on the order placed with CDC, updating as the situation changed in the field. This, by implication, required rapid communication, coordination, and assessment of needs and supplies across local health departments and health care facilities within their jurisdiction. As the shortage occurred at the optimal time for vaccination, the potential existed for heightened influenza activity in the coming flu season, requiring heightened surveillance for influenza activity as well as increased monitoring of health care resources, such as bed availability and Emergency Department traffic. There was an absolute and urgent need for statewide situational awareness by decision makers across all information flows related to the

Table 1. NY State Health Commerce System (HCS) Demographics and Usage as of July 2008

Organization or User Group	Number	Users
Clinical or Environmental Labs	1,329	5,455
Clinics and Treatment Centers	1,498	2,582
Home and Adult Care Facilities	1,658	9,749
Hospitals	230	14,665
Local Health Departments (LHD)	58	6,230
MDs and Practices	-	29,692
NYSDOH Central Health Office	1	3,942
NYSDOH Regional Health Offices	4	1,133
Nursing Homes	663	11,025
Other Clinical Practitioners and Practices	-	10,865
Pharmacies	1,113	3,173
Other Organizations	-	6,272
Usage Statistics		
250 applications		
6,500 User logins per day		
550,000 access hits per day		
10 Gbytes in transactions per day		

LHDs include NY City Department of Health and Mental Hygiene. Other clinical practitioners include dentists, veterinarians, nurse practitioners. Other organizations include schools, fire and EMS, federal and state agencies, tribal nations, managed care organizations, etc. Organization counts are by physical facility.

event. There was an equally critical need for rapid distribution of communiqués related to vaccine recommendations, allocation response plans, and alerts of local or statewide increases in influenza activity.

1.2 Informatics Framework and Information System Infrastructure Used in Response to Event

Over the past 13 years the NY State Department of Health (NYSDOH) has evolved both an informatics framework and a strategic information infrastructure to support information exchange with its health information trading partners. The infrastructure, Health Commerce System (HCS), is a secure, web-enabled portal supporting information exchange with all regulated health entities in NY [1,2,3,10]. The demographics and organizations using the HCS are shown in Table 1. The applications within the HCS support a broad range of health-related activities, from vital records and health care quality assurance and finance to disease registry and condition reporting, statewide communicable disease and laboratory reporting, arbovirus surveillance, child health insurance reporting, managed care, even prescription pad orders. The data and information flow within the HCS are shown in Figure 1. Given this mission, the HCS architecture is multi-tiered, highly available, and capable of full off-site disaster recovery. HCS is a platform well suited for response to public health emergencies, given its existing architecture and routine use by partner organizations involved in the response [1,2,3,10].

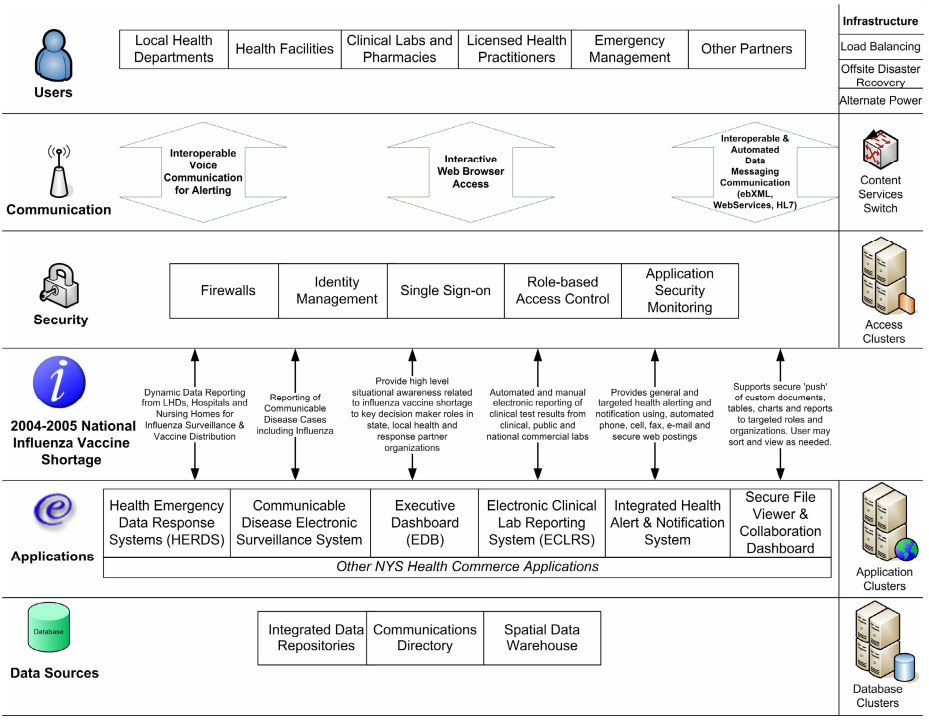


Fig. 1. NYS Health Commerce System Architecture and Functions Used in Response to the 2004 – 2005 National Influenza Vaccine Shortage

Thus an array of core PHEP information systems has evolved within the HCS to support health preparedness and response in NY. These are integrated with systems supporting core public health activities, such as statewide electronic disease and laboratory reporting. Representative PHEP systems are described in Table 2 (see also Figure 1). The HCS preparedness systems have supported statewide response to emergent infectious disease events, emergency disaster declarations, health resource shortages, elevated national threat levels, and high-profile security events [1,2,3]. The HCS infrastructure is also an integral component of NYSDOH incident management and PHEP plans [2].

2 Response to the Vaccine Shortage PHEP Event

2.1 Established Preparedness Systems and Surveillance in Place in Advance of the Event

The PHEP systems as listed in Table 2, or analogues thereof, were in operation in advance of the vaccine shortage and had been used in a number of preparedness drills and other PHEP events occurring before October 2004 [2]. The systems supporting the preparedness and response to the shortage included: Hospital Emergency

Table 2. Example Health Preparedness Functions and Systems Used in the NY State Health Commerce System

Generic Preparedness Function	Statewide Data System on NYSDOH HCS
Health Facility Surveillance, Dynamic Resource Reporting and Response	**HERDS (Hospital Instance)** Dynamic, multitasking system supporting live data feeds for electronic surveillance, surveys, resource and asset tracking, surge, bed availability, patient tracking. and emergency response in hospitals
Local Health Disease Reporting and Epidemiological Response	**CDESS** Statewide system for reporting of disease cases by LHDs; includes contact tracing and integration with ECLRS **HERDS (County Instance)** Dynamic, multitasking system supporting live data feeds for surveillance and survey distribution and return for LHDs
Alerting and Communication	**IHANS** Provides health alerting and notification using automated phone, cell, fax, e-mail, and web postings; integrated with ComDir **ComDir** Central directory of role and contact information for all HCS participants; maintained by coordinators at participant HCS organizations
Laboratory Reporting	**ECLRS** Supports automated and manual electronic reporting of clinical test results from clinical, public, and national commercial labs to CDESS
Data Visualization and Situational Awareness	**Executive Dashboard (EDB)** Access-limited system providing high-level situational awareness to key decision maker roles across response partner organizations; integrates multiple data streams (HERDS, CDESS) into graphical displays **Event WebSite** A we site within the HCS providing information postings and updates to entire HCS community during an event **Secure file viewer and collaboration dashboard system** Supports secure 'push' of static content to targeted roles and organizations; also used to support discussion forums between organizations and incident command involved in the event

HCS – NYSDOH Health Commerce System; **HERDS** – Health Emergency Response Data System; **CDESS** – Communicable Disease Electronic Surveillance System; **IHANS** – Integrated Health Alerting and Notification System; **ComDir** – Communications Directory; **ECLRS** – Electronic Clinical Laboratory Reporting System; **LHD** – Local Health Department

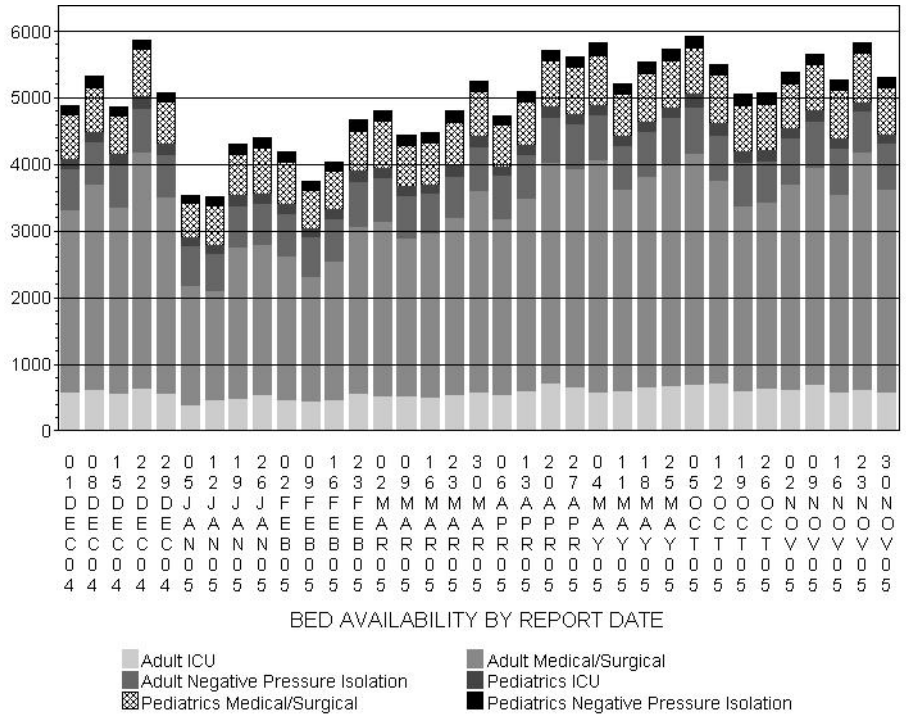


Fig. 2. Series of Statewide Aggregate of Hospital-Reported Bed Availability by Bed Type

Response Data System, or HERDS (see [1,2], Table 2), the Integrated Health Alerting and Notification System, or IHANS (see [1,2], Table 2) and the dashboard-secure collaboration form (see [1], Table 2). The HERDS system is currently deployed to all LHDs, nursing homes, adult- and home-care entities, and schools statewide. At the time of the vaccine shortage, HERDS was in routine use by all hospitals. HERDS prototypes were deployed to LHDs and nursing homes during the shortage. The interrelationship of these systems within the HCS is shown in Figure 1.

In response to elevated national threat levels in June 2003, NYSDOH used the HERDS system to implement ongoing statewide hospital bed availability and Emergency Department patient traffic reporting. The reporting system provided statewide, facility-specific surveillance of bed availability by specialty type (e.g., adult and pediatric, ICU, medical, surgical, burn, observational) and monitoring of ED patient admission traffic (e.g. see Figure 2). In response to the atypical severity and elevated pediatric hospitalizations and deaths associated with the 2003-2004 flu season, HERDS was used to establish an ongoing statewide hospital-based pediatric influenza surveillance system in early December 2003, one year prior to influenza becoming a nationally notifiable condition. The surveillance system reported facility-specific laboratory-confirmed cases of pediatric influenza admissions and deaths. An influenza vaccine inventory survey was also deployed in December 2003 to all hospi-tals. Bed availability reports during the 2003–2004 flu season corroborated the information received on elevated ED traffic or specialty bed utilization (e.g., ICU),

potential indirect surrogates to any general increase in influenza activity that year. As part of preparedness planning for the 2004 Republican National Convention in New York City, NYSDOH used the HERDS system to deploy a statewide survey of hospital critical assets in August 2004. The survey was an exhaustive inventory of currently staffed and surge-capacity beds by specialty: special treatment capacities (e.g., trauma and burn center, hyperbaric, decontamination); transportation capacities; durable and fixed equipment (e.g., adult and pediatric ventilators, cardiac monitors); personal protective equipment and pharmaceutical inventory; staff capacities by specialty; communication and generator capacities. The survey also included staff influenza vaccination rates. Much of this data is also essential to resource allocation and response to a local or large-scale influenza outbreak. Thus, as events turned out, in 2003 and 2004 NY State was using its PHEP systems to establish the surveillance and resource inventory activities that would be key in responding to the 2004 vaccine shortage.

2.2 Preparedness Response to the Event

To respond to the vaccine shortage, NY State needed to: 1) assess and update data on vaccine inventories, orders, and needs for priority risk groups in the state; 2) develop ordering requirements for CDC Phase II; 3) develop an in-state allocation and distribution plan based on up-to-the-minute data; 4) assure rapid and effective communication with LHDs and health facilities; 5) monitor the effects of increased influenza activity or hospital utilization due to vaccination shortfalls; 6) detect local increases in influenza activity; 7) assure overall situational awareness for NYSDOH executive incident command process and for external response partners. A detailed timeline of these activities is presented in Table 3.

2.2.1 Vaccine Assessment and Allocation Plan

Within 24 hours of the Chiron announcement, NYSDOH alerted its LHD and health care partners to the new vaccine priorities and provided guidance. Seven days later a complete statewide assessment of hospital vaccine needs was deployed and completed within 24 hours. Within 14 days NYSDOH decision makers had a complete picture of vaccine inventories, orders, and needs across hospitals and other health care facilities statewide. For facilities not using HERDS, gathering data was laborious and time consuming. However, by mid-November prototype HERDS instances were deployed to LHDs and nursing homes, allowing turnaround time similar to that for the hospital HERDS surveys. Within 9 days of the beginning of CDC Phase II, NYSDOH had developed a statewide, data-driven vaccine allocation plan and placed its vaccine orders with CDC. With the full electronic deployment of survey capability across facility types, NYSDOH was also able to update needs assessments continuously throughout November and December. In the case of the hospital HERDS reports, NY State was again able to initiate and turn around complete statewide surveys within 24 hours for the November and December vaccine updates.

2.2.2 Health Care Monitoring and Influenza Surveillance

The potential for ED overcrowding due to the impact of the vaccine shortage on influenza activity in the impending flu season was recognized early on. Facilities were alerted in October, well in advance of the onset of the flu season, as to the need to

Table 3. Timeline of Events, Actions, and Response Milestones during the National Influenza Vaccine Shortage, October-December 2004

Date	Public Health Action	Description/Purpose/Outcome
Ongoing since June 2003	Hospital bed availability and emergency department patient admission traffic surveillance	HERDS monitoring of hospital bed availability by specialty type and patients waiting in the Emergency Department. Situational awareness available to state, regional and local health offices as information is reported.
OCT 5	Influenza vaccine shipments from Chiron suspended	US vaccine supplies reduced by approximately one-half.
OCT 6	IHANS Health Alert/Notification	Official NYSDOH Commissioner Communiqué to LHDs and hospitals with new CDC vaccination guidelines and the rationale for vaccinating by priority groups.
OCT 12	Phase I CDC Plan (see [9])	CDC ships previously held orders; to be administered to high-risk groups.
OCT 12	First hospital vaccine needs assessment and inventory survey deployed statewide	HERDS survey on hospital vaccine and antiviral inventories, orders, and needs by risk (priority) group. 90% of hospitals (213/237) respond within 24 hours.
	IHANS Health Alert/Notification	Sent to all LHDs and hospitals; provided survey notification; vaccine distribution plan; treatment and infection control guidelines for respiratory outbreaks.
OCT 14	First vaccine survey of nursing homes, home healthcare, diagnostic and treatment centers	Surveys per hospital vaccine assessment. Reporting occurred via phone and file upload to HCS surveys as HERDS had not been deployed to these facilities.
OCT 19	NYSDOH Incident Command decision makers have full picture of vaccine needs	Situational awareness provided through Health Commerce System (HCS) data visualization system, based on data derived from hospital survey of OCT 12 and other facilities surveyed on OCT 14.
OCT 20	First update to situational awareness for external partners	Regional Health Offices, LHDs apprised of vaccine inventories and needs by risk group in their health facilities (based on data from OCT 14 surveys and HERDS).

Table 3. (*continued*)

OCT 26	Health Alert sent to all hospitals statewide	Official NYSDOH Commissioner Communiqué relaying potential for hospital overcrowding and guidance on ED preparedness and respiratory precautions.
NOV 5	Alert Notification on total vaccine shipped to date.	LHDS and Regional Health Offices notified that reports CDC Phase I vaccine shipments to facilities within their jurisdiction are available on the collaboration dashboard.
NOV 8	First HERDS-p LHD Vaccine Needs and Inventories Survey .	HERDS-p application prototype was deployed to LHDS for vaccine needs assessment. 95 % (54/57) of LHDs respond on deadline, within 24 hours.
NOV 9	Situational awareness update to external partners	LHDS and Regional Health Offices notified via IHANS that additional reports on CDC Phase I vaccine shipments are available on the collaboration dashboard system.
NOV 9	Second hospital vaccine needs and inventory survey	90% of hospitals statewide respond to HERDS survey on vaccine needs by priority risk group within 24 hours.
NOV 16	First HERDS-p vaccine survey of nursing homes	HERDS application prototype was deployed to nursing homes for vaccine needs assessment.
NOV 17	CDC Phase II vaccine allocation plan [9]	States able to place orders for vaccine for priority group needs unmet by CDC Phase I.
NOV 26	Vaccine allocation plan complete	Statewide vaccine allocation and distribution plan is complete and order placed. Plan based on all data gathered from LHDs, hospitals, nursing homes, adult care entities.
NOV 29	Alert notification to LHDs that hospital vaccine plan available	Official NYSDOH Commissioner Communiqué relays details of the hospital vaccine allocation plan and of related data available to LHDs on the collaboration dashboard system.
NOV 30	Alert notification to hospitals on vaccine allocation	IHANS notifies hospitals that vaccine shipment allocations/schedule are available on the collaboration dashboard system.

Table 3. (*continued*)

DEC 2	Alert notification to LHDs on vaccine allocation	Reports on vaccine doses allocated by LHD, in addition to the hospital allocation announced on NOV 30, are available on collaboration dashboard.
DEC 4	Enhanced influenza surveillance deployed to all hospitals in HERDS	Ongoing reporting of laboratory confirmed influenza related hospital admissions by age group. Situational awareness regarding these admissions available to state, regional, and local health offices as reported.
DEC 18 and DEC 20	Follow-up needs surveys for LHDs, nursing homes, and hospitals	NYSDOH Incident Command has access to data as reported; 90% of hospitals respond within 24 hours.
DEC 22	Situational awareness for external partners	Updated hospital and nursing home vaccine needs and allocations provided to LHDs through the collaboration dashboard system.

follow preparedness and HERDS surveillance procedures (Table 3). HERDS data reported by health facilities is accessible by State, regional, and local health departments as it is reported. Thus the HERDS bed availability survey, which tracked both bed utilization and ED patient traffic and had been in place since June 2003, allowed for detailed tracking and trending by aggregate, local, or facility-specific views during the 2004-2005 flu season (see Figure 2). This data stream, in combination with the baseline data of the HERDS critical asset survey (August 2004), allowed health officials at all levels to monitor, confirm, or rule out any indications or external (anecdotal) reports of ED or hospital overcrowding. At the statewide aggre-aggregate level, there was an increase in bed utilization (decreased availability) and ED traffic in early January 2004 (Figure 2).

In response to concerns over increased influenza activity, the HERDS-based pediatric hospital influenza surveillance initiated in the previous influenza season (2003-2004) was expanded to include laboratory-confirmed influenza admissions for all age groups and deployed in December 2004 (Table 3). Again, having the surveillance process in place from the previous season and a single common reporting interface (HERDS) greatly facilitated the reporting process and also provided an integral location for reviewing reporting streams. Notably, the increase in state-aggregated reports of hospital admissions for laboratory-confirmed flu in early January correlated with increased patients waiting in the ED and decreased bed availability (Figures 2, 3).

2.2.3 Situational Awareness

The IHANS alerting system served four roles in NY's efforts to provide situational awareness to external response partners: supporting advance preparedness; directly providing event-related content; notifying partners of the availability of new (or updates

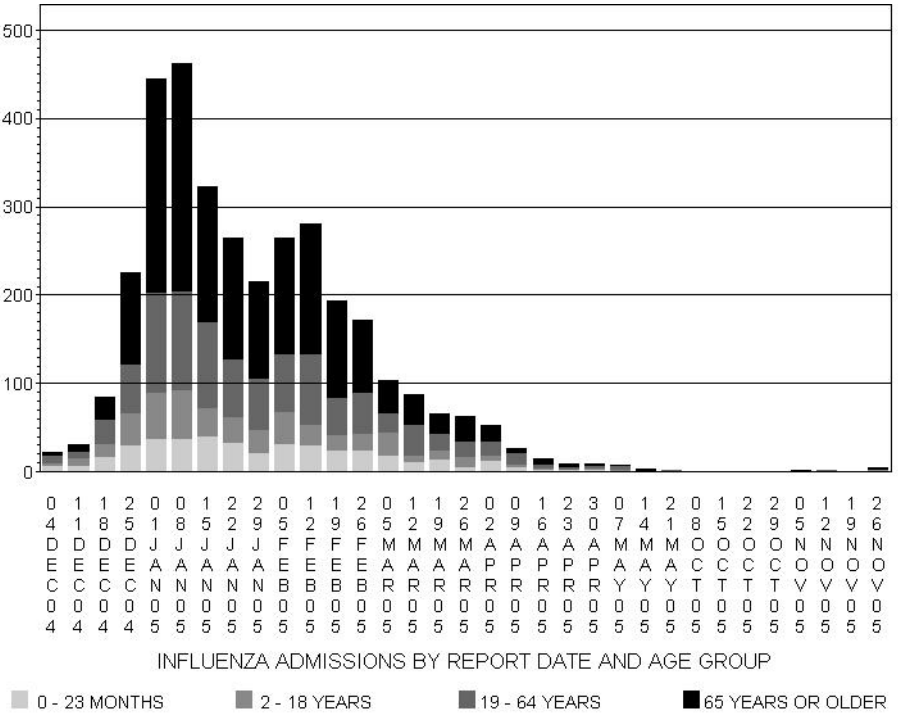


Fig. 3. Time Series of Statewide Aggregate Hospital Reports of Laboratory- Confirmed Influenza Hospital Admissions by Age Group

to) analytic products on the dashboard system; and notifying both organizations and health officials that HERDS surveys had been activated (see Table 3). The support of advance, or just-in-time, preparedness is illustrated by the Health Commissioner Alert of October 26, 2004 (Table 3). Advance recognition of the potential for increased ED overcrowding due to increased influenza activity led to a health alert being sent to health facilities to review ED overcrowding preparedness plans, institute respiratory precautions, and keep up to date with HERDS surveillance guidances.

In total, 28 health alerts related to influenza were sent using the IHANS system during October-December 2004. The topics ranged from vaccination recommendations to updates on the shortage, state and federal response plans, influenza activity updates, priority risk group recommendations, and collateral impacts of the shortage. The target audiences included LHDs as well as the health facilities (e.g., hospitals, nursing homes, individual providers) using the HCS system. Other alerts sent from IHANS (e.g., 2) were used to notify response partners as to the availability of data visualization products on the collaboration dashboard and the activation of surveys in HERDS.

Access to situational awareness data for external response partners was provided through reports, charts, and graphs derived from the HCS data visualization system and provided through the collaboration dashboard forum. These products provided information, in both aggregate and detail, integrated across the various data streams

and allocation plans related to the event (Table 3). The collaboration dashboard also allowed for electronic dialogue between state and local health departments regarding the data available on the dashboard. As shown in Table 3, throughout the event external response partners were able to access information related to vaccine needs across facility types, vaccine shipments, and allocations within jurisdiction. While all health organizations had access to HERDS data as it was reported, customized summaries of HERDS data feeds were also provided through the collaboration dashboard. HERDS data feeds are also available through the HCS GIS system, allowing spatial trending and drill-down access to facility-specific detail regarding surveillance data, bed availability, available assets, and vaccine needs.

3 Conclusions and Lessons Learned

The 2004–2005 vaccine shortage preceeded by the severe 2003–2004 flu season was exactly the type of PHEP event on which public health preparedness programs focus: "unexpected/without warning, national implications, widespread public anxiety and fear of illness and death" [10,11]. The presence of an established integrated informatics framework for health information exchange and PHEP in NY State conveyed significant advantages in advanced preparedness and just-in-time response to this health event. The key PHEP benefits of having this framework include: a demonstrable state of response readiness; rapid establishment and maintenance of situational awareness across response partners through just-in-time dynamic information-gathering activities; effective communication and coordination of a broad spectrum of response activities; rapid development and implementation of a data-driven response plan.

Lessons learned from the event response include the following:

- PHEP readiness is optimized when supportive information systems are embedded within an established, dual-use, informatics framework, such as the NYSDOH HCS system. HCS supports a broad-based community of health information trading partners who become response partners in a PHEP event. Among the many advantages of the system are economy of scale, familiarity with and trust of the system, common and standardized usability, depth and breadth of partner organization inclusion and communication, data integration, and new opportunities for synergies and linkage.
- Systems such as HERDS [2]—which support rapid, integrated, and flexible deployment of ongoing surveys across the universe of health care facilities and partners—are ideally suited to emergent PHEP events. Establishing the system as common practice through routine surveillance heightens both state and local ability to respond, as does gaining the routine participation of the many health care organizations whose help would be needed to respond to an emergency.
- A key lesson learned from the vaccine shortage was the advantage of involving all types of health organizations in preparedness systems. Early on in the event, key organization types (nursing homes, adult care facilities, and LHDs) had access to the HCS system but were not instantiated within the HERDS system. This resulted in the need for intensive out-of-band processing and manual work to capture and integrate

reports from these organizations. This was in contrast to hospitals, already established in HERDS, which were reporting on surveillance activities as well as turning around vaccine needs surveys within 24 hours. The scenario changed midway through the event, as new HERDS prototypes were deployed to LHDs and nursing homes. In response to its after-action analysis of the event, NYSDOH took two major steps to address the issue. First, it engaged a process to deploy the HERDS system to all regulated health facility organizations in NY, including rollout and training. Second, it created a new regulation requiring all health facility organizations in NY to maintain a cadre of skilled HCS users, maintain up-to-date contact information in the communications directory (Table 2), and use the HERDS system. As of this date all hospitals, nursing homes, adult and home care facilities, and schools utilize HERDS for surveillance and routine reporting. Other organizations in progress include clinics, clinical labs, and pharmacies.

Acknowledgments

We are indebted to Robert L. Burhans, Director of Health Preparedness, and Dr. Dale Morse, MD, Director of the Office of Science at NYSDOH, for their leadership and support. We also acknowledge the New York Association of County Health Officials, hospitals, and local health departments of NY State for their contributions to the successful response of New York State to public health events. This publication was supported by Cooperative Agreement Numbers U50/CCU223671 and U90/CCU216988 from the U.S. Centers for Disease Control and Prevention (CDC) and by Cooperative Agreement Number 6U3RHS05934 from the U.S. Health Resources and Services Administration (HRSA). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC or HRSA.

References

1. Gotham, I.J., Sottolano, D.L., Le, L.H., Primeau, M.J., Santilli, L.A., Johnson, G.S., Ostrowski, S.E., Hennessey, M.E.: Design and Performance of a Public Health Preparedness Informatics Framework: Evidence from an Exercise Simulating Influenza Outbreak. In: Castillo-Chavez, C., Chen, H., Lober, W., Thurmond, M., Zeng, D. (eds.) Infectious Disease Informatics and Biosurveillance: Research, Systems, and Case Studies, Springer, Heidelberg (Fall 2008) (in press)
2. Gotham, I., Sottolano, D., Hennessy, N., Napoli, J., Dobkins, G., Le, L.H., Burhans, R., Fage, B.: An integrated information system for all hazards health preparedness and response, New York State Health Emergency Response Data System (HERDS). Journal of Public Health Management and Practice 13(5), 487–497 (2007)
3. Gotham, I., Eidson, M., White, D., Wallace, B.J., Chang, H.G., Johnson, G.S., Napoli, J.P., Sottolano, D.L., Birkhead, G.S., Morse, D.L., Smith, P.F.: West Nile Virus: A Case Study in How NY State Health Information Infrastructure Facilitates Preparation and Response to Disease Outbreaks. Journal of Public Health Management Practice 7(5), 75–86 (2001)
4. US Centers for Disease Control and Prevention. Seasonal flu (last accessed August 25, 2008), `http://www.cdc.gov/flu/`

5. US Centers for Disease Control and Prevention. 2003–04 U.S. Influenza Season Summary (last accessed August 25, 2008), http://www.cdc.gov/flu/weekly/fluactivity.htm
6. US Centers for Disease Control and Prevention. 2004–05 U.S. Influenza Season Summary (last accessed August 25, 2008), http://www.cdc.gov/flu/weekly/fluactivity.htm
7. US Centers for Disease Control and Prevention. Update: Influenza Activity – United States, 2004–2005 Season. MMWR Weekly 54(13), 328–331 (2005)
8. US Centers for Disease Control and Prevention. Childhood Influenza Vaccination Coverage – Unites States, 2003–2004 Influenza Season. MMWR Weekly 55(04), 100–103 (2006)
9. US Government Accountability Office Influenza Vaccine. Shortages in 2004–05 Season Underscore Need for Better Preparation. GAO-05-984, 38 pages (2005)
10. Assn of State and Territorial Health Officials, The 2004/2005 Influenza Vaccine Shortage: Implications for Public Health Emergency Preparedness, Issue Report (January 2006)
11. Schoch-Spana, M., Fitzgerald, J., Kramer, B.R.: UPMC Influenza Task Force: Influenza Vaccine Scarcity 2004–2005: Implications for Biosecurity and Public Health Preparedness. Biosecurity and Bioterrorism: Biodefense Strategy, Practice and Science 3(3), 224–234 (2005)

Dynamic Network Model for Predicting Occurrences of Salmonella at Food Facilities

Purnamrita Sarkar, Lujie Chen, and Artur Dubrawski

The Auton Lab, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
http://www.autonlab.org

Abstract. Salmonella is among the most common food borne illnesses which may result from consumption of contaminated products. In this paper we model the co-occurrence data between USDA-controlled food processing establishments and various strains of Salmonella (serotypes) as a network which evolves over time. We apply a latent space model originally developed for dynamic analysis of social networks to predict the future link structure of the graph. Experimental results indicate predictive utility of analyzing establishments as a network of interconnected entities as opposed to modeling their risk independently of each other. The model can be used to predict occurrences of a particular strain of Salmonella in the future. That could potentially aid in proactive monitoring of establishments at risk, allowing for early intervention and mitigation of adverse consequences to public health.

Keywords: Link analysis, latent space models, social networks, food safety surveillance, risk based inspection.

1 Introduction

In the United States, about 40,000 cases are reported annually including 400 deaths from acute Salmonellosis (CDC, 2008). These statistics are observed despite widespread efforts of federal and local food safety offices aimed at mitigation of the involved risks to public health. Some of these efforts consider using routinely collected data to monitor and predict outcomes of microbial testing of food samples taken at processing facilities. The analytic techniques used so far (FSIS, 2008; Roure et al., 2007b; Roure et al., 2007a) focus on utilizing predictive models developed for the individual establishments under assumption of independence. In this paper, we evaluate utility of a network-based approach in which individual establishments are treated as entities in a network, interconnected via links corresponding to occurrences of specific serotypes of Salmonella observed at them.

Link structure learning algorithms have been proven useful in a variety of applications based on a social network paradigm (Madadhain & Smyth, 2005; Breiger et al., 1975). Recently, there were a few attempts to use them in the context of bio-security (Dubrawski et al., 2008; Reis et al., 2007). In this paper we

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 56–63, 2008.

apply DSNL (Sarkar & Moore, 2005), a latent space model originally developed for dynamic social network analysis, to predict occurrences of specific serotypes of Salmonella at food processing establishments.

In the proposed framework, the historical record of positive results of microbial tests for Salmonella, conducted at individual food production facilities, is used to construct a bipartite graph. Nodes of this graph represent respectively the individual facilities and the specific Salmonella serotypes. A facility node is linked with a serotype node if the particular type of Salmonella was isolated at the given plant during the period of observation. The hypothesis being tested is that the observed networking of facilities through serotype occurrences and can be predictive of the future microbial safety performance. To test it, we split the available historical data into independent training and test subsets in order to objectively measure prediction accuracies. We also compare our results with a baseline which only considers histories of individual establishments, independently of each other, to make predictions. Any gain of the predictive accuracy of the proposed approach over the baseline would indicate the potential utility of the network based approach in predicting the risk of future occurrences of microbial contamination in specific food processing establishments.

The details of the used real-world datasets, the experimental setup, and the obtained results are presented in a separate Section of the paper. The next Section briefly explains DSNL, the dynamic social network model used in the experiments.

2 Dynamic Social Network in Latent Space Model

In this section we briefly describe the network model and the algorithm used to train it (more details can be found in (Sarkar & Moore, 2005)). (Raftery et al., 2002) defined a latent space model which associates each entity with a location in a p-dimensional Euclidean space. The idea is that entities close in the latent space are more probable to form a link in the network. (Sarkar & Moore, 2005) generalized this static model to a Dynamic Social Network in Latent space model (DSNL). It is designed to explain changes of relationships between entities over time by allowing the latent coordinates to change smoothly over time. The main idea is to associate each link with a discrete timestep. The entities can move in the latent space between consecutive timesteps, but large moves are deemed improbable.

Let G_t be the graph observed at timestep t. The latent position of entity i at timestep t is denoted by the i^{th} row of the $n \times k$ positions matrix X_t. The model uses standard Markov assumption (Figure 1), similar to that widely used in Hidden Markov Models or Kalman Filters. It also involves two probability distributions. One generates G_t from X_t such that links between pairs of entities which are far away in the latent Euclidian space are less probable. The other distribution generates X_{t+1} from X_t, and controls the smoothness of transition. These distributions also aid in tractable learning of the maximum likelihood estimates of the latent positions of the entities. We want:

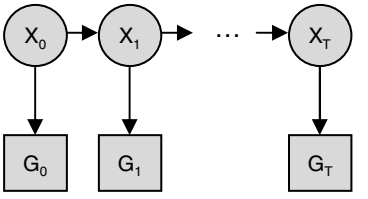


Fig. 1. Structure of the DSNL model

$$X_t = \arg\max_X P(X|G_t, X_{t-1}) = \arg\max_X P(G_t|X)P(X|X_{t-1}) \tag{1}$$

2.1 Model Description

The log-likelihood form of of the model equation (1) decomposes into two parts, i.e. $\log P(G_t|X)+\log P(X|X_{t-1})$. The first part, the observation model, measures how well the latent coordinates explain the observed graph. The second part, the transition model, penalizes large changes from the latent positions learned in the last timestep.

Denote the distance between entities i and j at timestep t as d_{ij}. Then, radius parameter r_i is introduced for each entity i. The value of this parameter can be learned from data. It captures the relative importance of an entity in the network. r_{ij} equals greater of the radii of entities i and j. The probability of a link between entities i and j is then estimated as:

$$p_{ij} = \frac{1}{1 + e^{(d_{ij} - r_{ij})}} \tag{2}$$

The probability that graph G_t was generated from coordinates X_t, that is the observation model, is therefore given by the following:

$$p(G_t|X_t) = \prod_{i \sim j} p_{ij} \prod_{i \not\sim j} (1 - p_{ij}) \tag{3}$$

Apparently, it is possible to eliminate quadratic computation of the observation model over all pairs of links by introducing a biquadratic kernel. As a result of this simplification, two entities have high probability of linkage only if their latent coordinates are within the radius r_{ij} of one another. Beyond this range there is a constant noise probability of linkage.

The transition model used is simply Gaussian:

$$X_t \sim \mathcal{N}(X_{t-1}, \sigma^2)$$

The parameter σ controls the smoothness of transition, that is large values of σ allow large changes of latent coordinates of the entities from one timestep to the next.

2.2 Algorithm Description

The optimization algorithm has two-phases. First, the latent coordinates are initialized by a time-dependent variation of the classical multidimensional scaling (Borg & Groenen, 1997). The solution combines the evidence from the current observation, and the last timestep's locations. These estimates are then used to initialize the non-linear optimization.

Classical multidimensional scaling (MDS) takes as input an $n \times n$ matrix of non-negative distances D where D_{ij} denotes the target distance between entity i and entity j. It produces an $n \times p$ matrix X where the i^{th} row is the position of entity i in p-dimensional latent space. Let the coordinates of n points in a p dimensional Euclidean space be given by $x_i, (i = 1 : n)$ where $x_i = (x_{i1}, ..., x_{ip})$. Without delving into much detail, MDS transforms the pairwise distance matrix D into a similarity matrix $\tilde{D}$ using linear transformations. The solution is given by the following formula in which F denotes Frobenius' metric:

$$X = \arg\min_X |\tilde{D} - XX^T|_F \tag{4}$$

In order to incorporate temporal smoothness into this objective (Sarkar & Moore, 2005) proposed to minimize the following objective function:

$$X_t = \arg\min_X |\tilde{D}_t - XX^T|_F + \lambda |XX^T - X_{t-1}X_{t-1}^T|_F \tag{5}$$

The first part of it is identical to the standard MDS objective. The second part encourages small changes in pairwise distances between two consecutive timesteps. The parameter λ controls the relative importance of the past and present evidence. The above optimization problem has a closed form solution:

$$X_t X_t^T = \frac{1}{1+\lambda}\tilde{D}_t + \frac{\lambda}{1+\lambda} X_{t-1}X_{t-1}^T \tag{6}$$

X_t can be obtained via eigen-decomposition of the right hand side of (6). It is possible to compute X_t using an iterative solver in $O(n^2 f + n + pn)$ time per iteration, where n is the number of entities, p is the number of latent dimensions, and f is the fraction of non-zero entries in the underlying matrix.

The solution (6) becomes the starting point of a nonlinear optimization for the next time step using conjugate gradient. Due to the use of the biquadratic kernel, the computation of gradient of the likelihood only needs to consider entities that lie within one another's radius. This eliminates the need for iterating over all pairs of entities and the computations can be executed efficiently using KD-trees (Preparata & Shamos, 1985) in $O(rn + n \log n)$ time, where r is the average number of in-radius neighbors.

3 Experiments

The objective of the experiments summarized below is to evaluate the utility of exploiting the structure of connectivity between food establishments and strains

(serotypes) of Salmonella, in predicting occurrences of a particular strain of Salmonella in the future. The assumption being made states that if the two (or more) establishments share a historical pattern of co-occurring strain-specific isolates, we should expect them to be linked in a similar way in the near future, provided that the environmental drivers of the underlying processes remain stationary. If the assumption holds, the proposed approach could become a useful part of the risk-prediction toolkit in the food safety domain.

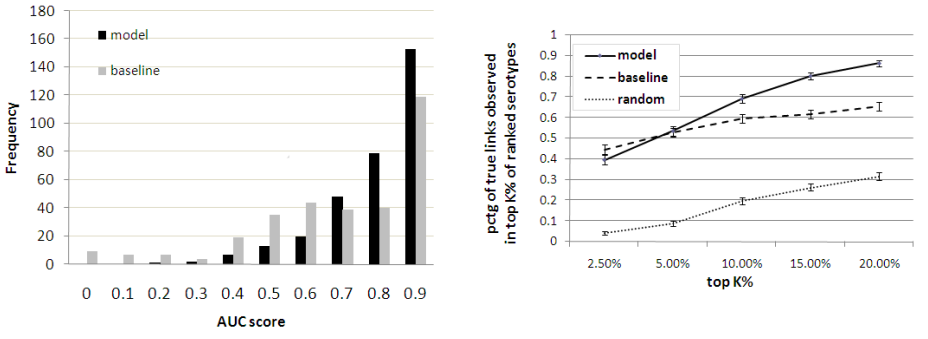
To evaluate the performance of the proposed method we chose a baseline algorithm which predicts two entities to share a link if the weighted average of their co-occurrences over the past timesteps is high. In this baseline model we exponentially down-weigh the evidence from older data. Note that the baseline only looks at the individual establishment's history of isolates, and does not model transitivity of similarity like th network model does.

The data used in our experiments is an excerpt of the record of regulatory sampling of food for Salmonella conducted at a subset of USDA regulated establishments from January 2005 till December 2007. Each record in this data represents a positive result of a microbial test of a sample of food taken at a specific establishment. The data includes the information of the specific serotype of the isolated Salmonella. It consists of over 7,000 records of positive tests involving about 750 unique establishments and over 90 unique serotypes.

The training data is used to build bipartite graphs of connections between establishments and serotypes. If a specific serotype was observed at a specific establishment during a period of observation at least once, these two entities will be linked in the graph. The network model is trained using data arranged in two observation periods corresponding to years 2005 and 2006. The trained model is then used to predict links between establishments and serotypes over the period of 2007. The predictions are then compared with the actual observations recorded in the test data. On the average in each year the data included around 400 unique establishments, 70 serotype entities, and $1,000$ links between the establishments and the serotypes.

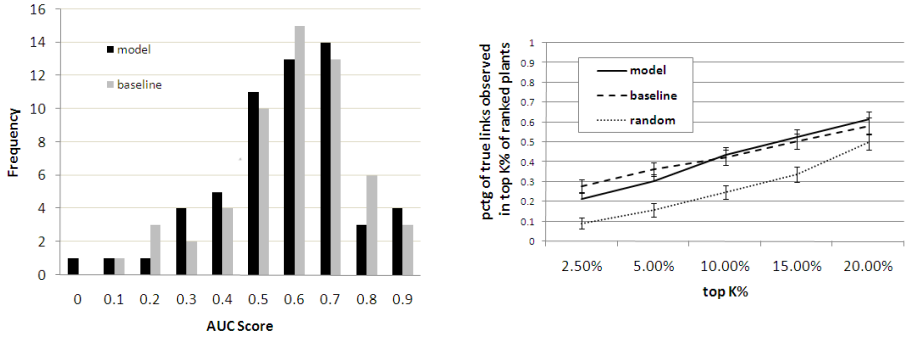
Consider the following scenario: the analyst has a historical record of Salmonella positives isolated at various establishments during two or more past periods of observation (years). Now for any given establishment she can use the trained network model to predict the top $k\%$ most likely serotypes which might be observed at that establishment during the subsequent period of time. A symmetrical question would be: for a given serotype of interest, recommend the $k\%$ establishments which are most likely to record such isolate during the next period of time. Predictive utility of the proposed method in answering such questions can be easily quantified in terms of the AUC (Area Under the ROC Curve) scores and recall scores obtained by comparing the probabilities estimated using the model and the actual links observed in the test set.

In order to address the first of the two questions, for each establishment we rank the predicted probabilities of occurrence of every serotype according to DSNL model and, separately, according to the baseline model. From these rankings we can compute: (1) AUC scores; and (2) Fractions of the true links between



(A) (B)

Fig. 2. A. Distribution of AUC scores and B. Average recall values for the top k% of ranked serotypes for establishments.



(A) (B)

Fig. 3. A. Distribution of AUC scores and B. Average recall values for the top k% of ranked establishments for serotypes.

the establishment and serotypes among the top k% of the serotypes ranking lists produced by the respective models. The AUC scores measure the overall ability of a model to produce rankings aligned with the actual observations recorded in the test data. The recall scores provide a similar indication for the alignment of the highest ranking predictions. Such scores may make more sense than AUCs in practical situations whenever constraints on investigative resources limit the analysts to inspecting only a few top findings.

Figure 2A depicts the distribution of AUC scores computed for the individual establishments using DSNL and the baseline model. The DSNL distribution is more skewed to the right: it leads to a better overall predictive performance with respect to numerous establishments when compared to the baseline.

Figure 2B shows the recall scores computed for $k = 2.5\%$, 5%, 10%, 15% and 20%, respectively for the DSNL model, the baseline and the random predictor

(note that the random predictor's expected AUC score equals 0.5 and therefore it was omitted from the graph in Figure 2A). It is clear that for values of k greater than 5% the proposed algorithm outperforms the plausible baseline.

Executing a pairwise t-test for comparing the scores of DSNL and the baseline at individual establishments leads to significant p-values. For the AUC results, the p-value equals $6.6 \cdot 10^{-16}$. The p-values obtained for recall scores were 0.4, $2.1 \cdot 10^{-05}$, $2.1 \cdot 10^{-14}$, and $9.6 \cdot 10^{-21}$, respectively for k set to 5%, 10%, 15% and 20%. Note that as suggested in the graph, only the differences in performance between DSNL and the baseline model for $k \leq 5\%$ is insignificant.

We have performed a similar analysis for the other scenario (predicting rankings of establishments most likely to record a specific isolate of Salmonella). Figure 3A and 3B demonstrate the distributions of AUC scores and the recall results for this task. The plots indicate that neither the baseline nor DSNL have performed very well. The p-values for pairwise t-test also indicate that the difference between the performances of the model and the baseline is not statistically significant. Apparently, from the perspective of a serotype, the basic hypothesis of network-based similarity models does not hold as strongly as for the processing establishments. Two establishments may appear similar in their performance if they operate in similar food-processing environments, and intuitively that may to lead to similar patterns of results of microbial testing. However, a similar claim cannot be made as strongly for a pair of serotypes. This is probably why the model and the baseline outperform the random process by a large margin in the first task (serotype prediction), but not in the second (establishment prediction).

4 Conclusion and Future Work

We prsented an application of a Dynamic Social Network in Latent space model (DSNL) to prediction tasks in the food safety surveillance domain. The proposed approach exploits similarities in the historical records of positive isolates of Salmonella serotypes at different food production facilities. In that, it expands on the previous work which typically assumes independence of the processes governing microbiological performance of the individual facilities. Conducted experiments indicate predictive utility of modeling the system as a network of food establishments interconnected via specific Salmonella serotypes. We compared the network model with a simple, but usually hard-to-beat baseline algorithm which does assume independence. We examined two link prediction tasks using real data. In one of these tasks, the proposed network model significantly outperforms the baseline. In the other task, however, neither of them performed well.

In this paper, we focused on evaluating predictive performance of the DSNL model. Howver, it can also be used as a powerful visualization tool to help understanding of the evolution of the underlying network over time. In addition, we have limited our attention to a fenomenological concept of links in the network. We have not considered perhaps more intuitive ways in which the facilities may be interrelated e.g. due to shared supply channels or due to common

corporate memberships. The observed predictive utility of our network is considerably interesting on its own, nonetheless we plan to expand the framework by incorporating those more intuitive connectivity patterns in order to measure their effect on predictive power of the attainable models. On the algorithm level, we plan to investigate how much improvement in prediction performance can be achieved by weighting the links (i.e. by the frequncy of co-occurrences) and we will verify the impact of varying temporal granularity of obsrvation and prediction periods on the predictive performance of the model.

References

Borg, I., Groenen, P.: Modern multidimensional scaling. Springer, Heidelberg (1997)

Breiger, R.L., Boorman, S.A., Arabie, P.: An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. J. of Math. Psych. 12, 328–383 (1975)

CDC. Centers for Disease Control and Prevention: General information on Salmonellosis (2008), `http://www.cdc.gov/ncidod/dbmd/diseaseinfo/salmonellosisg.htm`

Dubrawski, A., Chen, L., Ostlund, J.: Using AFDL algorithm to estimate risk of positive outcomes of microbial tests at food establishments. Advances in Disease Surveillance 5 (2008)

FSIS. Food Safety Inspection Service, U.S. Department of Agriculture: Data analysis for public health risk-based inspection system for processing and slaughter. Appendix E - Data analyses (2008), `http://www.fsis.usda.gov/oppde/nacmpi/feb2008/processingappendix-e-041808.pdf`

Madadhain, J., Smyth, P.: Learning predictive models for link formation. In: Sunbelt 25: International Sunbelt Social Network Conference (2005)

Preparata, F., Shamos, M.: Computational geometry: An introduction. Springer, Heidelberg (1985)

Raftery, A.E., Handcock, M.S., Hoff, P.D.: Latent space approaches to social network analysis. J. Amer. Stat. Assoc. 15, 460 (2002)

Reis, B.Y., Kohane, I.S., Mandl, K.D.: An Epidemiological Network Model for Disease Outbreak Detection. PLoS Med. (2007)

Roure, J., Dubrawski, A., Schneider, J.: Learning specific detectors of adverse events in multivariate time series. Advances in Disease Surveillance 4 (2007a)

Roure, J., Dubrawski, A., Schneider, J.: A study into detection of bioevents in multiple streams of surveillance data. In: Zeng, D., Gotham, I., Komatsu, K., Lynch, C., Thurmond, M., Madigan, D., Lober, B., Kvach, J., Chen, H. (eds.) BioSurveillance 2007. LNCS, vol. 4506, pp. 124–133. Springer, Heidelberg (2007b)

Sarkar, P., Moore, A.: Dynamic social network analysis using latent space models. In: Proceedings of the 19th Annual Conference on Neural Information Processing Systems (NIPS) (2005)

Network-Based Analysis of Beijing SARS Data

Xiaolong Zheng[1], Daniel Zeng[1,2], Aaron Sun[2], Yuan Luo[1], Quanyi Wang[3], and Feiyue Wang[1]

[1] The Key Lab of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China
[2] Department of Management Information Systems, The University of Arizona, USA
[3] Beijing Center for Disease Control and Prevention, China

Abstract. In this paper, we analyze Beijing SARS data using methods developed from the complex network analysis literature. Three kinds of SARS-related networks were constructed and analyzed, including the patient contact network, the weighted location (district) network, and the weighted occupation network. We demonstrate that a network-based data analysis framework can help evaluate various control strategies. For instance, in the case of SARS, a general randomized immunization control strategy may not be effective. Instead, a strategy that focuses on nodes (e.g., patients, locations, or occupations) with high degree and strength may lead to more effective outbreak control and management.

Keywords: SARS, Complex network analysis, Weighted networks.

1 Introduction

Severe Acute Respiratory Syndrome (SARS) was first found in the Guangdong Province of China in November, 2002. During its 2003 outbreak, 8,098 confirmed cases were reported in more than 30 countries within a very short period of time [1]. Among them, 2,521 cases were reported in Beijing, representing close to one third of the entire world-wide infected population.

The SARS outbreak prompted a world-wide public health response and has had a dramatic impact on the Chinese public health system as to infectious disease prevention, outbreak detection, and response. From a research perspective, significant efforts from both public health and related fields including but not limited to various subareas of informatics and computer-based modeling, have been devoted to studying the evolution and transmission patterns of SARS for future prevention and treatment purposes.

The SARS literature from the perspective of infectious disease informatics has also been growing [1, 2, 3, 4, 5, 6]. For example, several control measures have been proposed to control the outbreak of the SARS epidemic [3, 4]. Spatial analysis of SARS cases has been explored recently to reveal the associations between various related epidemic determinants [1]. Some authors have developed network-based mathematical models to analyze the transmission patterns of the SARS outbreak and to predict the outbreak diversity [2, 6]. Despite the significance and importance of using real-world SARS data to validate and evaluate

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 64–73, 2008.

these modeling efforts, however, very limited work has been done from an empirical analysis perspective, partially due to the difficulty in accessing pertinent epidemiological data.

Our research aims to bridge some of the existing gaps in the empirical analysis line of work and to better connect the complex network analysis literature with infectious disease informatics practice. In this reported research, we used the Beijing SARS data provided by the Beijing Center for Disease Control. By modeling patients, locations (districts in Beijing), and patient occupations as nodes, respectively, and treating contacts or infections as edges, we have constructed and analyzed three kinds of SARS-related networks: the patient contact network, the district network, and the occupation network. In Section 2, we provide a brief introduction to the data and the network-based analysis methods used in our research. Section 3 presents findings based on the patient contact network. Sections 4 and 5 report findings based on the weighted district and occupation networks, respectively. We conclude the paper in Section 6 by discussing ongoing and future research.

2 Data and Analysis Methods

Our Beijing SARS data were collected from an extensive survey of 624 confirmed SARS patients from 14 administrative districts in Beijing, covering the period from March 10, 2003 to May 13, 2003. These patients worked in 21 categories of occupations. We have followed previous studies (e.g., [2, 6]) to define an "infectious link" pointing from patient A to patient B, if it is highly likely that A transmitted the SARS virus to B. In total, 447 such infectious links were identified.

In our analysis, we first constructed a patient contact network based on infectious links as typical in existing epidemiological studies, and analyzed this network. However, with SARS being a unique and highly contagious airborne epidemic disease, personal contacts uncovered in the patient surveys or interviews alone may not provide sufficient information to fully explain the transmission patterns. As such, in an exploratory attempt, we also constructed two additional networks: a location/district network and an occupation network to further illustrate the spreading of the SARS epidemic in various parts of Beijing and among different occupational categories. We study these two networks as "weighted networks," with the weight w_{ij} defined over an directed edge from node i to j given as the total number of the infectious links from i to j. Further, we study node "strength" s_i defined as $s_i = \sum_{j=1}^{N} w_{ij}$ for node i, where N is the total number of nodes in the network [7]. This strength measure can be indicative of the ability to spread the disease from a given node.

In the next section, we investigate the topological properties of the patient contact network and its evolution pattern. We then discuss findings based on the weighted district and occupation networks.

3 Patient Contact Network

Patient contact networks can provide useful information concerning disease transmission and have been studied in the existing literature in various scenarios [8, 9, 10]. Using the Beijing SARS data described in Section 2, we first study the degree distribution of the SARS contact network and then investigate its temporal evolution.

3.1 Degree Distribution

It is well-known in the complex network analysis literature that network representations of a large number of real systems can be characterized by a node degree distribution with a power-law tail [11]. This is of particular importance in epidemiology since in this case the expected reproductive number may be unbounded [12]. In epidemiology, the reproductive number is defined as the number of secondary infections generated by one patient. This concept plays a key role in understanding the dynamic process of epidemics and in evaluating impact of control measures on the spread of infection [13]. Fig. 1 shows that the SARS contact network also follows a power-law distribution. The blue line corresponds to a power-law tail $P(k) \sim k^{-\gamma}$with $\gamma = 2.8076$. Not surprisingly, this result shows that the SARS infectious network is a scale-free network, with the implication that the expected reproductive number can be unbounded. A public health implication of this finding is that the traditional disease control approach based on random immunization (which has been shown to be effective in many epidemic outbreaks [8]) may not be effective (unless, of course, the entire population can be treated), because untreated hubs, albeit small in number, can still lead to rapid and large-scale infections [8]. Instead, an alternative control method targeting at containing highly connected nodes can be much more effective.

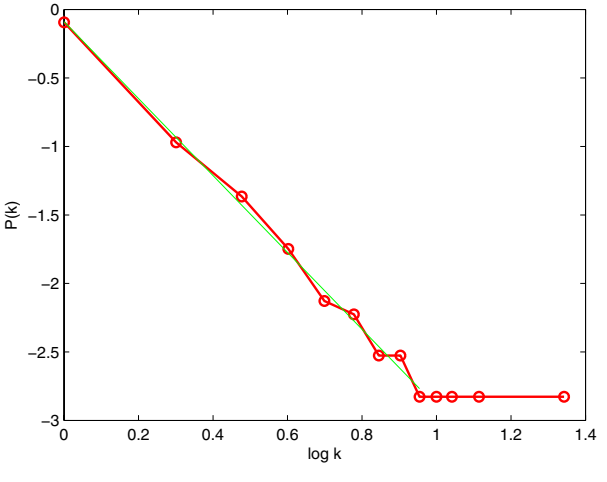


Fig. 1. Degree distribution of SARS patient contact network

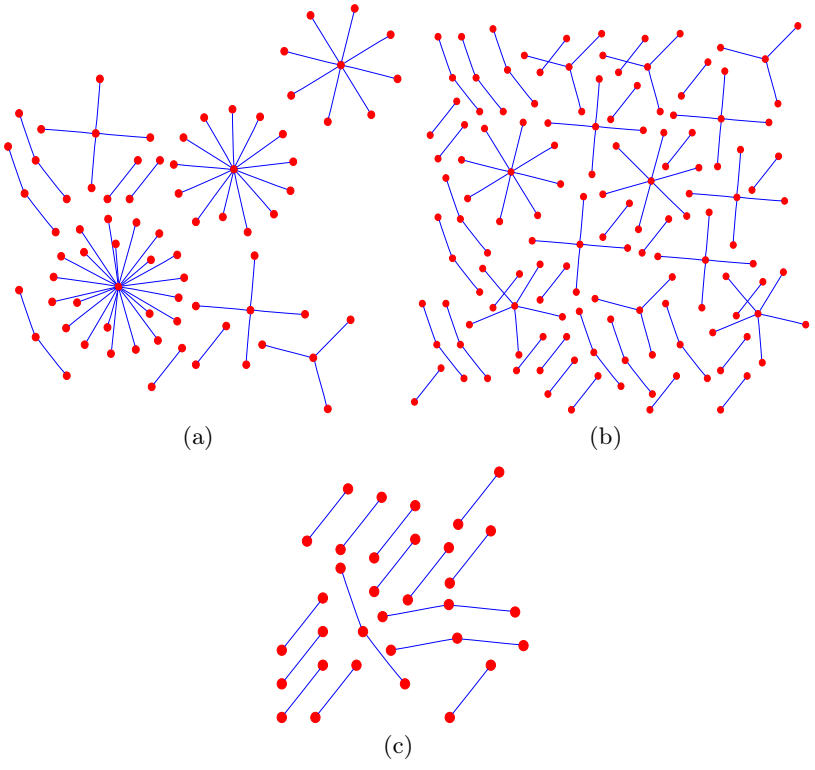


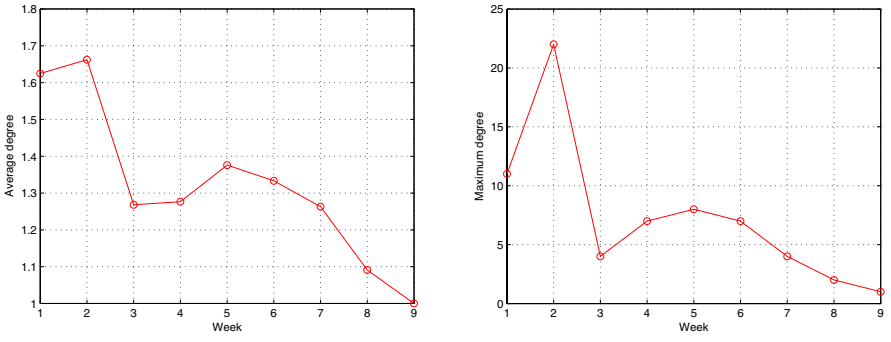
Fig. 2. SARS patient contact network in (a)–Week 2, (b)–Week 5, and (c)–Week 8

3.2 Network Evolution

It has been pointed out that in different phases, the transmission of epidemics may exhibit different patterns [2, 12]. Since the records in our nine-week SARS dataset are timestamped, we are able to observe the evolution of the SARS patient contact network over time. We plot three weekly "snapshots" of the contact network in Fig. 2.

The node degree in the contact network can be used to measure the node's disease spreading ability [2]. Fig. 3 (a) and (b) plot the average and maximum degrees for 9 consecutive weeks, respectively. We notice that in the first two weeks, the contact network has a relatively high average and maximum degree. These measures start to decrease with time after Week 5. This decrease can be attributed (at least partially) to several strong control measures implemented by the government after April 14, 2003 (which is in Week 5).

A connected component of the contact network is a set of nodes in which each node is connected to at least one edge. Connected components can be used to demonstrate the extent to which an epidemic can spread within a population [14]. We define the component ratio as the number of connected components divided by the total number of nodes in the network. From Fig. 4, we observe



(a) (b)

Fig. 3. Network evolution. (a)–average degree and (b)–maximum degree.

that in the first two weeks, the ratio is relatively small. After some fluctuations in the next two weeks, it starts to increase gradually. Part of this observation is due to the fact that, during the first few weeks, most SARS patients were misdiagnosed as having tuberculosis [15]. The isolation and quarantine controls were not enforced for these patients during this initial period of outbreak. After April 14, 2008, as strong control measures were taken, the epidemic was brought under control.

4 Weighted District Network

The patient contact network analyzed in the previous section can provide insights as to SARS transmission patterns among patients. However, for SARS, personal contact information available does not provide sufficient explanation for the underlying transmission patterns of this epidemic (partially due to the incomplete nature of contact information acquired through patient surveys or interviews). Geographical information is also crucial to gain a better understanding of the epidemic [16].

Fig. 5 plots the weighted district network (WDN). We analyzed the infection transmission patterns through the WDN. The results on the cumulative weight distribution are shown in Fig. 6 (a). As we can observe, the cumulative weight distribution follows a strongly right-skewed distribution, indicating a high degree of heterogeneity in the WDN.

To better understand the WDN, we define $s_d(k_d)$ as the average strength of nodes with degree k_d. Theoretically, if $s_d(k_d)$ and k_d are uncorrelated, then $s_d(k_d) \sim k_d^{\alpha}$ with $\alpha = 1$. In this case, weights cannot provide any additional information than degrees [17]. Our analysis shows that the observed $s_d(k_d)$ increases with k_d as $s_d(k_d) \sim k_d^{\alpha}$ with the exponent $\alpha = 1.8775$. The findings are plotted in Fig. 6 (b). Table 1 displays the top five district strengths. These results indicate that the strengths of nodes are strongly correlated to degrees in the WDN.

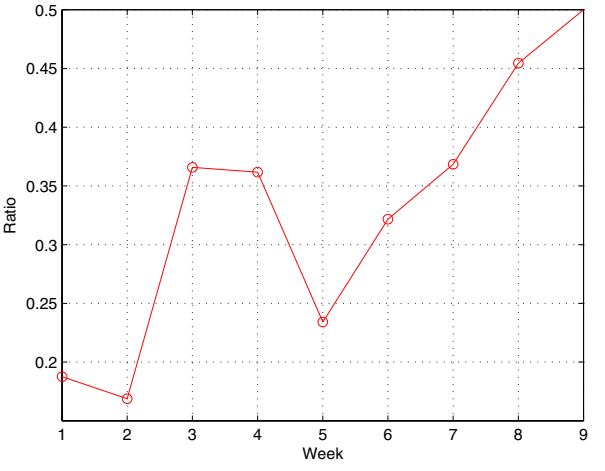


Fig. 4. Changes in network component ratio over time

Table 1. Top five district (node) strengths

District	Chaoyang	Haidian	Dongcheng	Fengtai	Changping
Strength	241	152	113	97	87

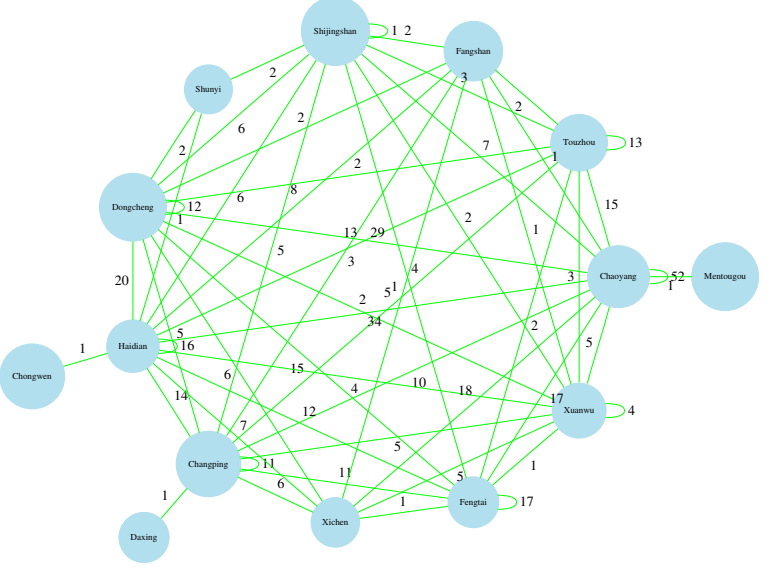
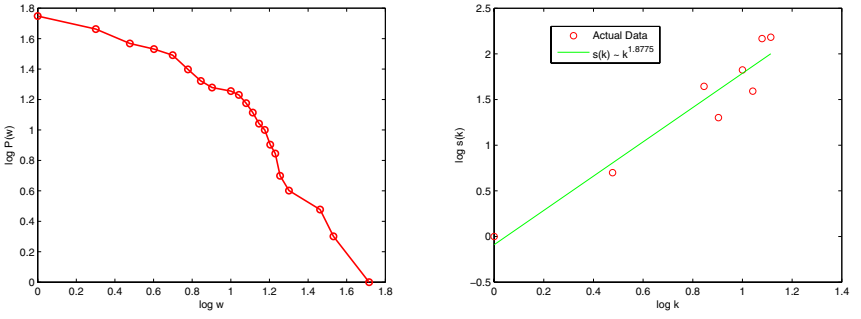


Fig. 5. Weighted district network



(a) (b)

Fig. 6. (a) WDN cumulative weight distribution. (b) Average node strength s_d as a function of node degree k_d.

One possible explanation is that both Chaoyang and Haidian District are major financial districts with (combined) more than 15 million permanent and temporary residents. Individuals in such densely-populated areas are more likely to be exposed to the epidemic and further spread the disease.

5 Weighted Occupation Network

Disease transmissions often happen in workplaces and in turn occupations can have an impact on the spreading patterns of infectious diseases [18]. In this section, we analyze the SARS transmission patterns based on the weighted occupation network (WON) as shown in Fig. 7. Our preliminary results show that the WON can reveal some additional insights.

The cumulative weight distribution of the WON, shown in Fig. 8 (a), follows a right-skewed distribution. Table 2 lists these five occupations with top strengths. The retiree category has the maximum strength 153, while the strengths of the other four occupations, military personnel, governmental employees, unemployed, and industry workers have relatively smaller occupation strengths.

This analysis shows that not all the occupations have equal probabilities to be infected with the SARS virus. For instance, the retiree population was more susceptible to be infected because of their lowered immune function. In the Chinese society, the retirees play an active role in family functions and child care and their working sons and daughters are in different occupations. Previous papers (e.g., [15]) have also reached similar conclusions. From a outbreak control perspective, those occupations with strong strengths need to be closely monitored.

Following an analysis procedure similar to that used for the WDN, we conclude that for the WON the average node strength s_d increases with the degree k_d and that $s_d(k_d) \sim k_d^{\beta}$, with the exponent $\beta = 1.6142$, which is larger than 1. This result is shown in Fig. 8(b), indicating that node strength is also strongly correlated to degree in the WON.

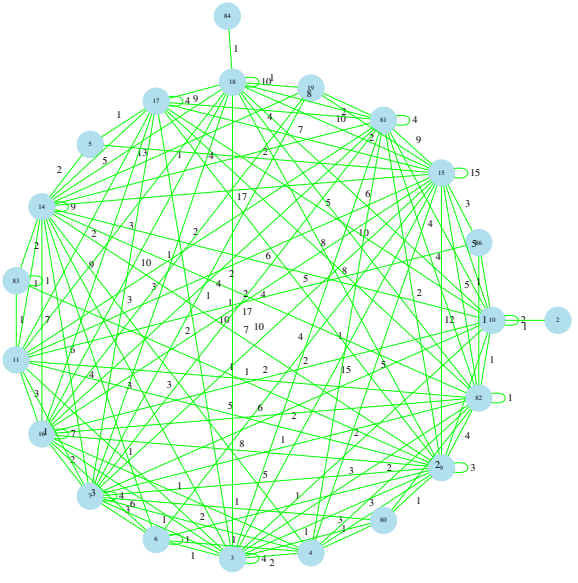
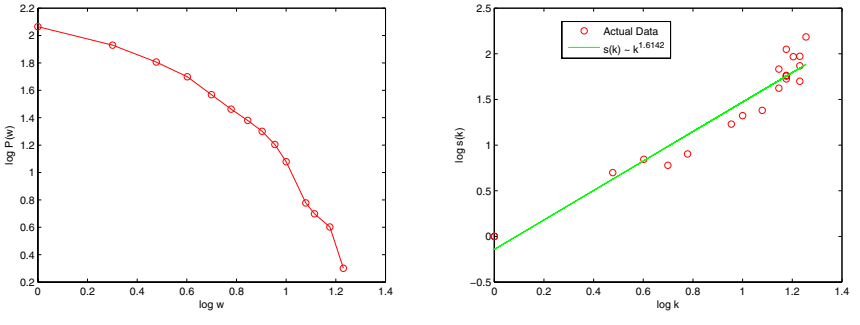


Fig. 7. Weighted occupation network

Table 2. Top five occupation (node) strengths

Occupation	Retiree	Military	Government Employee	Unemployed	Industry
Strength	153	112	94	93	74



(a) (b)

Fig. 8. (a) WON cumulative weight distribution. (b) Average node strength s_d as a function of node degree k_d.

6 Concluding Remarks

In this paper, we analyze Beijing SARS data from a complex network analysis perspective. To better understand the SARS epidemic transmission patterns and evolution, we have studied three networks derived from the patient survey data, including a patient contact network, a weighted district network, and a weighted occupation network. The patient contact network possesses the scale-free degree distribution and its temporal evolution (as measured by average degree, maximum degree, and component ratio) exhibits some interesting patterns that can be explained by various control measures implemented during the SARS outbreak in Beijing. In both weighted district and occupation networks, the weights follow right-skewed distributions and the strengths of nodes are strongly correlated to their degrees. These observations and analysis results indicate that the traditional random isolation control method may not be effective. Instead, a more effective control program should target at nodes with high degree and strength.

Due to various difficulties in data collection, the Beijing SARS dataset used in our study may not be complete in that some infectious links may be missing. Our current work focuses on inferring some of these missing links for analysis purposes using methods similar to those reported in [19, 20, 21]. We are also working on analyzing various topological and distributional properties of weighted networks. The results are expected to benefit epidemiological data analysis in general.

Acknowledgments

We would like to thank Pin Yan, Zhidong Cao, Fen Xia, Huiqian Li, Su Li, Cheng Nie, Hao Lu, Changli Zhang, and Xiaoli Wu for useful discussions and helpful suggestions.This work is supported in part by NSF #IIS-0839990 and #IIS-0428241; NNSFC #60621001; MOST #2006CB705500 and #2006AA010106; and CAS #2F05N01 and #2F07C01.

References

[1] Wang, J.F., Christakos, G., Han, W., Meng, B.: Data-driven exploration of 'spatial pattern-time process-driving forces' associations of SARS epidemic in Beijing, China. Journal of Public Health, 1–11 (2008)
[2] Meyers, L.A., Pourbohloul, B., Newman, M.E.J., Skowronski, D.M., Brunham, R.C.: Network theory and SARS: predicting outbreak diversity. Journal of Theoretical Biology 232, 71–81 (2005)
[3] Riley, S., et al.: Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science 300, 1961–1966 (2003)
[4] Dye, C., Gay, N.: Modeling the SARS Epidemic. Science 300, 1884–1885 (2003)
[5] Becker, N.G., Glass, K., Li, Z., Aldis, G.K.: Controlling emerging infectious diseases like SARS. Mathematical Biosciences 193, 205–221 (2005)
[6] Small, M., Tse, C.K.: Clustering model for transmission of the SARS virus: application to epidemic control and risk assessment. Physica A: Statistical Mechanics and its Applications 351, 499–511 (2005)

[7] Barrat, A., Barthélemy, M., PastorSatorras, R., Vespignani, A.: The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America 101, 3747–3752 (2004)
[8] May, R.M., Lloyd, A.L.: Infection dynamics on scale-free networks. Physical Review E 64, 066112 (2001)
[9] Newman, M.E.J.: Spread of epidemic disease on networks. Physical Review E 66, 016128 (2002)
[10] Pastor-Satorras, R., Vespignani, A.: Epidemic Spreading in Scale-Free Networks. Physical Review Letters 86, 3200 (2001)
[11] Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47 (2002)
[12] Vazquez, A.: Causal tree of disease transmission and the spreading of infectious diseases. DIMACS Series in Discrete Mathematics and Theoretical Computer Science 70, 163–179 (2006)
[13] Haydon, D.T., Chase-Topping, M., Shaw, D.J., Matthews, L., Friar, J.K., Wilesmith, J., Woolhouse, M.E.J.: The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. Proceedings of Royal Society London B 270 (2003)
[14] Liljeros, F., Edling, C.R., Amaral, L.A.N.: Sexual networks: implications for the transmission of sexually transmitted infections. Microbes and Infection 5, 189–196 (2003)
[15] Shen, Z., Ning, F., Zhou, W., He, X., Liu, C., Chin, D.P., Zhu, Z., Schuchat, A.: Superspreading SARS Events, Beijing, 2003. Emerging Infectious Diseases 10, 256–260 (2004)
[16] Chen, Y., Tseng, C., King, C., Wu, T.J., Chen, H.: Incorporating Geographical Contacts into Social Network Analysis for Contact Tracing in Epidemiology: A Study on Taiwan SARS Data. In: Zeng, D., Gotham, I., Komatsu, K., Lynch, C., Thurmond, M., Madigan, D., Lober, B., Kvach, J., Chen, H. (eds.) BioSurveillance 2007. LNCS, vol. 4506, pp. 23–36. Springer, Heidelberg (2007)
[17] Bagler, G.: Analysis of the airport network of India as a complex weighted network. Physica A: Statistical Mechanics and its Applications 387, 2972–2980 (2008)
[18] Vazquez, A.: Epidemic outbreaks on structured populations. Journal of Theoretical Biology 245, 125–129 (2007)
[19] Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature 453, 98–101 (2008)
[20] Hao, M., Irwin, K., Michael, R.L.: Effective missing data prediction for collaborative filtering. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, Amsterdam (2007)
[21] Ronald, K.P.: The problem of disguised missing data. SIGKDD Explorations Newsletters 8, 83–92 (2006)

Tutte Polynomials and Topological Quantum Algorithms in Social Network Analysis for Epidemiology, Bio-surveillance and Bio-security

Mario Vélez[1], Juan Ospina[1], and Doracelly Hincapié[2]

[1] Logic and Computation Group
Physical Engineering Program
School of Sciences and Humanities
EAFIT University
Medellin, Colombia
{mvelez,jospina}@eafit.edu.co
[2] Epidemiology Group
National School of Public Health
University of Antioquia
Medellin, Colombia
doracely@guajiros.udea.edu.co

Abstract. The Tutte polynomial and the Aharonov-Arab-Ebal-Landau algorithm are applied to Social Network Analysis (SNA) for Epidemiology, Biosurveillance and Biosecurity. We use the methods of Algebraic Computational SNA and of Topological Quantum Computation. The Tutte polynomial is used to describe both the evolution of a social network as the reduced network when some nodes are deleted in an original network and the basic reproductive number for a spatial model with bi-networks, borders and memories. We obtain explicit equations that relate evaluations of the Tutte polynomial with epidemiological parameters such as infectiousness, diffusivity and percolation. We claim, finally, that future topological quantum computers will be very important tools in Epidemiology and that the representation of social networks as ribbon graphs will permit the full application of the Bollobás-Riordan-Tutte polynomial with all its combinatorial universality to be epidemiologically relevant.

Keywords: Social Network Analysis, Tutte Polynomial, Aharonov-Arab-E bal-Landau algorithm, Topological Quantum Computation, Basic Reproductive Number, Borders.

1 Introduction

Human factors behind public health security for handling emerging diseases (SARS, Ebola, etc.) and re-emerging diseases (cholera, tuberculosis, influenza), and for avoiding threats of bioterrorist attacks, the geo-ecologic global crisis and the global shortage of food and fuel among others, were analyzed by The World Health Organization in the report entitled, "A safer future: global public health security in the 21st century" [1].

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 74–84, 2008.

Close and continuous networking groups at local and global levels have been observed in recent decades, with serious implications for the spread of infectious diseases [2].

In epidemiology in general and in biosurveillance and biosecurity in particular, the social network analysis (SNA) have been relevant in understanding these complex social phenomena [3].

In SNA, the mathematical theory of graphs and networks has a prominent role, since the social networks are represented by graphs. In algebraic graph theory, graphs and networks are characterized using graph polynomials such as the chromatic polynomial, the Tutte polynomial and the Bollobás-Riordan-Tutte polynomial [4].

Graph polynomials encode information about a graph and useful information about topological and geometrical properties of a graph may be extracted combinatorially from the algebraic structure of the graph polynomial. For example, the Tutte polynomial is able to encode a huge amount of information about the topological properties of a graph and hence the Tutte polynomial is relevant for epidemiology.

In Computational SNA applied to epidemiology, two lines of development have been recently explored: the Numerical Computational SNA (NCSNA) and the Algebraic Computational SNA (ACSNA). In NCSNA, numerical measurements for complex social networks are computed [3]. In ACSNA, algebraic objects like polynomials are computed for complex social networks. Another field is the application of topological Quantum Computation (TQC) in epidemiology, and specifically the possible application of the topological quantum algorithms for the Tutte polynomial [5].

This paper explores potential applications of the ACSNA and the TQC in epidemiology. Applications of Tutte polynomial and Topological Quantum Computers in Epidemiology, Bio-Surveillance and Bio-Security are described.

2 Tutte Polynomial in Epidemiology

In the Numerical Computational Social Network Analysis many packages are used to extract numerical and graphical information from a given social network. An example is illustrated in Figure 1. This figure shows a maplet which is able to obtain two numerical measurements named “stratum” and “compactness” from the displayed network [6]. The maplet is also able to obtain the adjacency matrix of the given network.

On the other side, in the Algebraic Computational Social Network Analysis, every network is characterized, not directly by a numerical measurement, but by an algebraic object such as a polynomial that codifies the combinatorial properties of such a social network. An example is the Tutte polynomial, which is the standard universal invariant topological polynomial for planar graphs.

Social networks are usually represented by standard planar graphs for which it is possible to define the Tutte Polynomial. For example, for the graph on the right hand side of Figure 1 the corresponding Tutte polynomial is given by

$$T(x, y) = x^2 (y + x + x^2) (x^5 + 4 x^4 + 6 x^3 + 4 x^3 y + 4 x^2 + 9 x^2 y + 3 x^2 y^2 + x + 7 x y^2 + 2 x y^3 + 6 x y + y + 3 y^3 + y^4 + 3 y^2) .$$

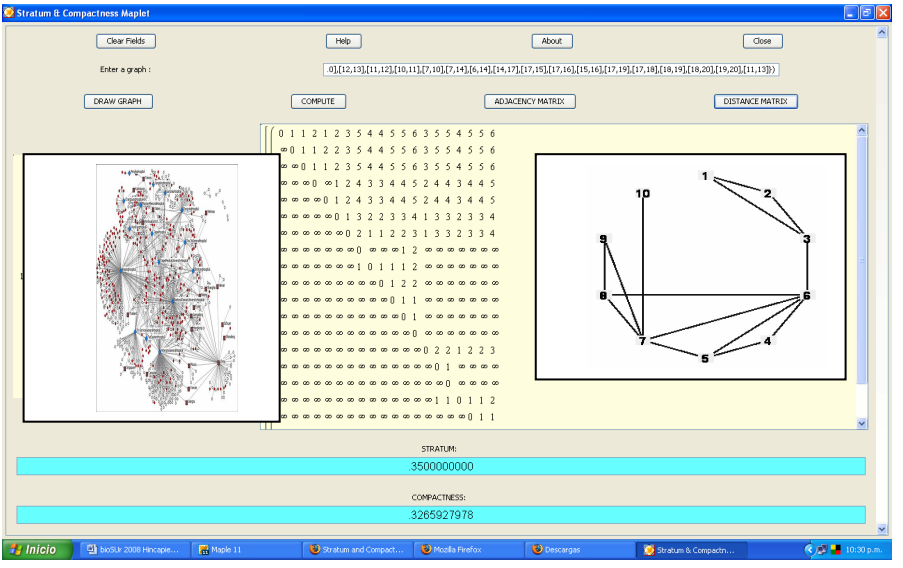


Fig. 1. A Maplet in Numerical Computacional Social Network Analysis

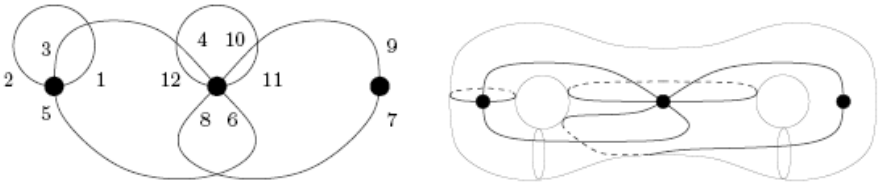


Fig. 2. A ribbon graph and its embedding in a bi-torus

In biosurveillance and bio security, social networks are more adequately represented using the so-called oriented ribbon graphs, which are graphs embedded in oriented surfaces. For the ribbon graphs, the Bollobás-Riordan-Tutte polynomial is defined and such a polynomial is a three-variable polynomial that generalizes the Tutte polynomial. An example of a ribbon graph is given below in the Figure 2 [7].

The ribbon graph in Figure 2 has only three vertices and six edges. In Figure 2 such a ribbon graph is depicted as embedded in a Riemann surface with genus two and is bi-torus. For this ribbon graph, the corresponding Bollobás-Riordan-Tutte polynomial is given by [7]

$$C(\mathbb{G}) = Z^2Y^4 + 2XZY^3 + 4ZY^3 + X^2Y^2 + 3XY^2 + +3XZY^2 + 4ZY^2 + 2Y^2 + 2X^2Y + 6XY + 4Y + X^2 + 2X + 1.$$

2.1 Network Evolution Described Via Tutte Polynomials

The social networks relevant in Epidemiology, Bio-Surveillance and Bio-Security are dynamic networks with a topological structure which is changing through time. An

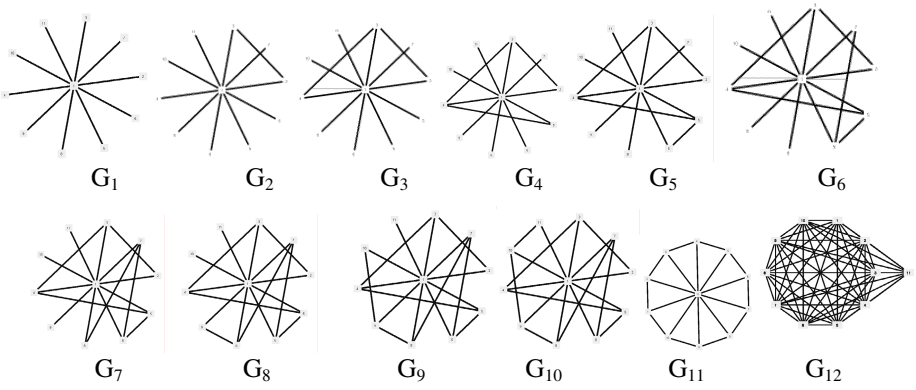


Fig. 3. Example of Random Network Evolution

example is depicted in the Figure 3. This figure shows the evolution of a network from an initial configuration denoted G_1 to a final configuration G_{12}. The sequence of graphs G_1, G_2, …,G_{12} can be represented as a sequence of the corresponding Tutte polynomials for the graphs G_1, G_2, …,G_{12}.

The evolution of the network is now viewed as a transition from the Tutte Polynomial $T(G_1,x,y)$ to the Tutte Polynomial $T(G_{12},x,y)$. Explicitly, the Tutte Polynomials for the graphs G_1, G_2,…, G_5 are given by

$$T(G_1, x, y) = x^{10}$$

$$T(G_2, x, y) = x^8 (y + x + x^2)$$

$$T(G_3, x, y) = x^7 (y^2 + y + x + 2xy + 2x^2 + x^3)$$

$$T(G_4, x, y) = x^6 (y^3 + 2y^2 + 4xy + 2xy^2 + 3x^2 + y + x + 3x^2 y + 3x^3 + x^4)$$

$$T(G_5, x, y) = x^5 (y^4 + 3y^3 + 7xy^2 + 2xy^3 + 9x^2 y + 3y^2 + 6xy + 4x^2 + 3x^2 y^2 + 6x^3 + y + x + 4x^3 y + 4x^4 + x^5)$$

From the Tutte polynomials corresponding to the social networks depicted in Figure 3, we may obtain numerical measurements about the connectivity of the contact networks. An example is shown in Table 1.

Every column in Table 1 is an increasing index of the complexity of the network. For example, epidemiologically, T(G,1,1) may be interpreted as the probability of infection for members of the social network G. This table shows that the probability of infection in the social network G_3 is eight times greater than the probability of infection in the social network G_1.

Table 1. Numerical Evaluations of the Tutte polynomial for networks in Figure3

Graph	T(G, 1,1)	T(G, 2, 1)	T(G, 1,2)	T(G, 2,0)	T(G,3,3)
G_1	1	1024	1	1024	59049
G_2	3	1792	4	1536	98415
G_3	8	3072	14	2304	170586
G_4	21	5248	48	3456	299619
G_5	55	8960	164	5184	528525
G_6	144	15296	560	7776	933606
G_7	377	26112	1912	11664	1649889
G_8	987	44576	6528	17496	2916135
G_9	2584	76096	22288	26244	5154426
G_{10}	6765	129904	76096	39366	9110859
G_{11}	15125	215230	215230	59046	17580753
G_{12}	2357947691	4767440679	35641657548953344	39916800	17 * 1022

2.2 Analysis of Reduced Networks Using Tutte Polynomials

Figure 4a shows a social network with personal and geographical contacts for which some nodes represent patients, other nodes represent hospitals and other different nodes represent geographical areas. These networks with geographical contacts are relevant in the analysis of pandemics such as SARS [3], avian flu, etc.

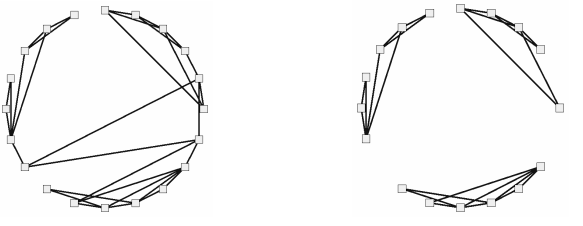


Fig. 4. A Network and its Reduced Network where some nodes are deleted

For the Network in Figure 4a, the corresponding Tutte polynomial is given by

$$\begin{aligned} T(G, x, y) = {} & x\,(y^2 + y + 2\,x\,y + x + 2\,x^2 + x^3)\,(y^4 + 4\,y^3 + 6\,y^2 + 3\,y + 10\,x\,y + x\,y^3 \\ & + 6\,x\,y^2 + 3\,x + 8\,x^2\,y + 2\,x^2\,y^2 + 7\,x^2 + 3\,x^3\,y + 7\,x^3 + 4\,x^4 + x^5)\,(x^6 + 5\,x^5 + 5\,x^4\,y \\ & + 10\,x^4 + 16\,x^3\,y + 10\,x^3 + 4\,x^3\,y^2 + 14\,x^2\,y^2 + 18\,x^2\,y + 5\,x^2 + 4\,x^2\,y^3 + 14\,x\,y^2 \\ & + 8\,x\,y + x + 9\,x\,y^3 + 3\,y^4\,x + 5\,y^3 + 4\,y^2 + y + 3\,y^4 + y^5)\,(y + x + x^2)^2 \end{aligned}$$

Figure 4b shows a reduced social network resulting from the social network depicted in Figure 4a, when three nodes are removed. The corresponding Tutte polynomial for this reduced social network takes the form

$$\begin{aligned} T(G_r, x, y) = {} & (y + x + x^2)\,(y^3 + 2\,y^2 + y + 2\,x\,y^2 + 4\,x\,y + x + 3\,x^2\,y + 3\,x^2 + 3\,x^3 + x^4)\,(\\ & x^5 + 4\,x^4 + 6\,x^3 + 4\,x^3\,y + 9\,x^2\,y + 3\,x^2\,y^2 + 4\,x^2 + 6\,x\,y^2 + 6\,x\,y + x + 3\,x\,y^3 + y \\ & + 2\,y^3 + y^4 + 3\,y^2)\,(y^2 + y + 2\,x\,y + x + 2\,x^2 + x^3) \end{aligned}$$

An algebraic measurement of the difference between the original network and its reduced network can be obtained as a relative difference between the Tutte polynomials for both networks

$$\frac{\mathrm{T}(G, x, y) - \mathrm{T}(G_r, x, y)}{\mathrm{T}(G, x, y)} = (129\,x^{11} - 4\,y^6 + 18\,x^6\,y^6 - 9\,x^2\,y^3 + 704\,y^4\,x^3 - 32\,x\,y^3 + x^{14}$$
$$+ 1039\,x^8\,y + 45\,x\,y^7 + 67\,x^9\,y^3 + 10\,x^{13} + 4\,x^9\,y^4 + 5\,x^{10}\,y^3 - 18\,y^4\,x - 46\,y\,x^3$$
$$+ 56\,x^4\,y^7 - 15\,x\,y^2 + 9\,x^5\,y^7 + 6\,x^{11}\,y^2 + y^{10}\,x + 190\,x^7 + 4\,x^4\,y^8 - 52\,x^2\,y^2 + 46\,x^{12}$$

$$+ 714\,x^6\,y + 352\,x^8\,y^3 + 967\,x^8\,y^2 + 245\,x^{10} - 17\,y\,x^2 - 10\,x^5 + 308\,x^8 - 7\,x^3$$
$$+ 76\,x^{11}\,y + 1743\,x^6\,y^2 + 1696\,x^6\,y^3 + 555\,x^5\,y^5 + 1347\,x^5\,y^4 - y^7 - 8\,y^5 - 6\,y\,x^4$$
$$+ 9\,x^{12}\,y - 9\,y^4 - 5\,y^3 - y^2 + 23\,y^8\,x + 291\,x^7\,y^4 + 35\,x^7\,y^5 + 328\,x^9 + 3\,x^3\,y^2$$
$$+ 412\,x^4\,y^2 + 344\,x^3\,y^3 + 23\,x\,y^5 + 51\,y^6\,x + 679\,x^9\,y + 272\,x^2\,y^5 - 2\,y\,x - x^2$$

$$+ 383\,y^6\,x^3 + 1133\,x^4\,y^3 + 150\,x^2\,y^4 + 7\,y^9\,x + 239\,y^6\,x^2 + 37\,x^2\,y^8 + 124\,x^2\,y^7$$
$$+ 683\,x^3\,y^5 + 54\,x^8\,y^4 - 18\,x^4 + 75\,x^{10}\,y^2 + 293\,x^{10}\,y + 59\,x^6 + 362\,x^9\,y^2 + x^8\,y^5$$
$$+ 124\,x^3\,y^7 + 253\,x^5\,y + 5\,x^2\,y^9 + 20\,x^3\,y^8 + 1173\,x^5\,y^2 + 1799\,x^5\,y^3 + 828\,x^4\,y^5$$
$$+ 1311\,x^4\,y^4 + 300\,x^4\,y^6 + 118\,x^5\,y^6 + 1072\,x^7\,y + 1615\,x^7\,y^2 + 994\,x^7\,y^3 + 203\,x^6\,y^5$$

$$+ 822\,x^6\,y^4 + y^9\,x^3) \,/\, ((y + x + x^2)\,x\,(y^4 + 4\,y^3 + 6\,y^2 + 3\,y + 10\,y\,x + x\,y^3 + 6\,x\,y^2$$
$$+ 3\,x + 8\,y\,x^2 + 2\,x^2\,y^2 + 7\,x^2 + 3\,y\,x^3 + 7\,x^3 + 4\,x^4 + x^5)\,(x^6 + 5\,x^5 + 5\,y\,x^4 + 10\,x^4$$
$$+ 16\,y\,x^3 + 10\,x^3 + 4\,x^3\,y^2 + 14\,x^2\,y^2 + 18\,y\,x^2 + 5\,x^2 + 4\,x^2\,y^3 + 14\,x\,y^2 + 8\,y\,x + x$$
$$+ 9\,x\,y^3 + 3\,y^4\,x + 5\,y^3 + 4\,y^2 + y + 3\,y^4 + y^5))$$

Numerical evaluations of the algebraic index of lost connectivity are:

$$\frac{\mathrm{T}(G, 1, 1) - \mathrm{T}(G_r, 1, 1)}{\mathrm{T}(G, 1, 1)} = 0.9593810445, \quad \frac{\mathrm{T}(G, 2, 1) - \mathrm{T}(G_r, 2, 1)}{\mathrm{T}(G, 2, 1)} = 0.9942398100,$$

$$\frac{\mathrm{T}(G, 1, 2) - \mathrm{T}(G_r, 1, 2)}{\mathrm{T}(G, 1, 2)} = 0.9806076277, \quad \frac{\mathrm{T}(G, 2, 0) - \mathrm{T}(G_r, 2, 0)}{\mathrm{T}(G, 2, 0)} = 0.9919354839$$

The numerical evaluations show that the reduced network in Figure 4b only keeps 5% of the original connectivity in the network of the Figure 4a.

2.3 Spatial Models with Bi-networks, Memories and Borders

Figure 5 shows a case where a country is considered as a circle with open boundaries, every sub-region is viewed as a complex network of localities and each locality is represented by a social network. Both the considered Bi-Networks and the open borders are spatial effects which are complemented by a temporal effect such as memory. The spatial propagation of a disease from the borders to the interior of the country, incorporating the effects of Bi-Networks and Memories, is examined here. Specifically, we will

present the explicit form of the basic reproductive number for our spatial model. We assume that human diffusion is spatially inhomogeneous and the corresponding diffusivity is a function of the spatial coordinates.

The SIR model spatially extended with a specified network of geographical localities and with a given contact network for every locality was assumed. The forces of infection and removal are affected by random noise which is described using exponentially decaying memory functions. As will be seen, the basic reproductive number is determined by the combinatorial properties of the networks, such as the degree distribution and the Tutte polynomial and its generalizations.

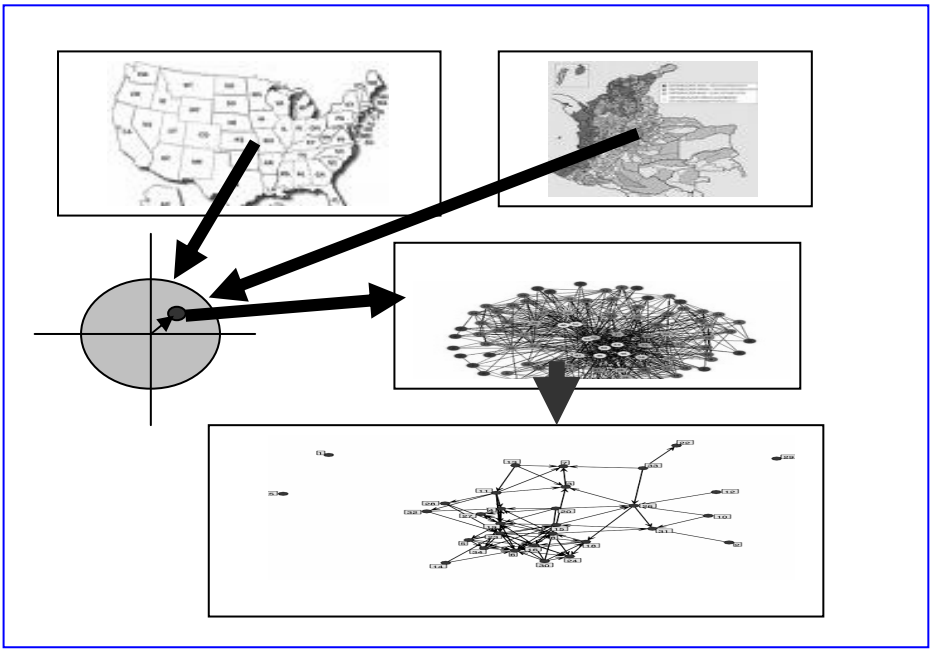


Fig. 5. Spatial Epidemic Model with Bi-Networks, Memories and Borders

The equations of the model are:

$$\frac{\partial}{\partial t} Y_i(r,t) = \frac{\frac{\partial}{\partial r}\left(\eta\, r^{(\lambda+1)}\left(\frac{\partial}{\partial r} Y_i(r,t)\right)\right)}{r}$$

$$+ \int_0^t \frac{\mathbf{e}^{(-\varepsilon_0(t-\tau))}\left(\frac{\partial}{\partial r}\left(\eta_1\, r^{(\lambda+1)}\left(\frac{\partial}{\partial r} Y_i(r,\tau)\right)\right)\right)}{r} d\tau + \beta\, N_i\, Y_i(r,t)$$

$$+\beta N_i\left(\left(\sum_{j=1}^{k} v_{j,i}\, Y_j(r,t)\right)-v_{i,i}\, Y_i(r,t)\right)+\beta N_i\left(\left(\sum_{j=1}^{k} v_{i,j}\, Y_j(r,t)\right)-v_{i,i}\, Y_i(r,t)\right)$$

$$-\gamma\, Y_i(r,t)+\beta_1 N_i\int_0^t \mathbf{e}^{(-\varepsilon_1(t-\tau))}\, Y_i(r,\tau)\, d\tau$$

$$+\beta_1 N_i\int_0^t \mathbf{e}^{(-\varepsilon_3(t-\tau))}\left(\left(\sum_{j=1}^{k} v_{j,i}\, Y_j(r,\tau)\right)-v_{i,i}\, Y_i(r,\tau)\right)d\tau$$

$$+\beta_1 N_i\int_0^t \mathbf{e}^{(-\varepsilon_4(t-\tau))}\left(\left(\sum_{j=1}^{k} v_{i,j}\, Y_j(r,\tau)\right)-v_{i,i}\, Y_i(r,\tau)\right)d\tau$$

$$-\gamma_1\int_0^t \mathbf{e}^{(-\varepsilon_2(t-\tau))}\, Y_i(r,\tau)\, d\tau$$

,

where $Y_i(r,t)$ is the density of infected individuals in the locality *i* at time t and coordinate *r*. The infectiousness parameters are denotedβ, the removal forces are denotedγ, and the parameters of memory are denoted. The human mobility between localities is measured by $v_{i,j}$, the human diffusivity is denoted with a parameter of spatial heterogeneity denoted. N_i is the total number of individuals in the locality i.

Using a standard method [8], the following form for the basic reproductive number is derived

$$R_{0,k,M,1,G,0}=\frac{N\beta\left(1+2\nu(k-1)+\dfrac{\beta_1}{\beta\,\varepsilon_1}+\dfrac{\nu(k-1)\,\beta_1\left(\dfrac{1}{\varepsilon_3}+\dfrac{1}{\varepsilon_4}\right)}{\beta}\right)E_G(d^2)}{\gamma k\left(\dfrac{5.784025\left(\dfrac{\delta_1}{\varepsilon_0}+\delta\right)}{\gamma a^2}+\dfrac{\gamma_1}{\gamma\,\varepsilon_2}+1\right)E_G(d)}$$

where $E_G(d)$ is the mean degree for the contact network G and $E_G(d^2)$ is the second moment for the degree of the contact network.

3 Topological Quantum Computers in Epidemiology

Computation of Tutte polynomials is a #P-hard problem [9]. This problem cannot be confronted using standard Turing machines and the corresponding classical computers.

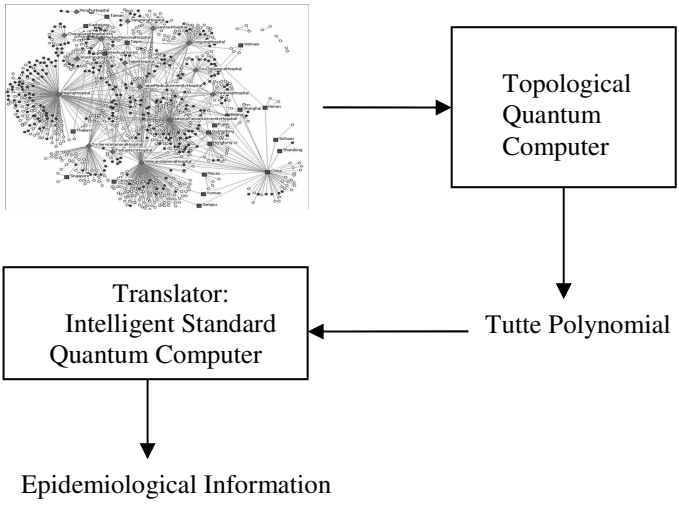


Fig. 6. Possible future application of Topological Quantum Computer in Epidemiology

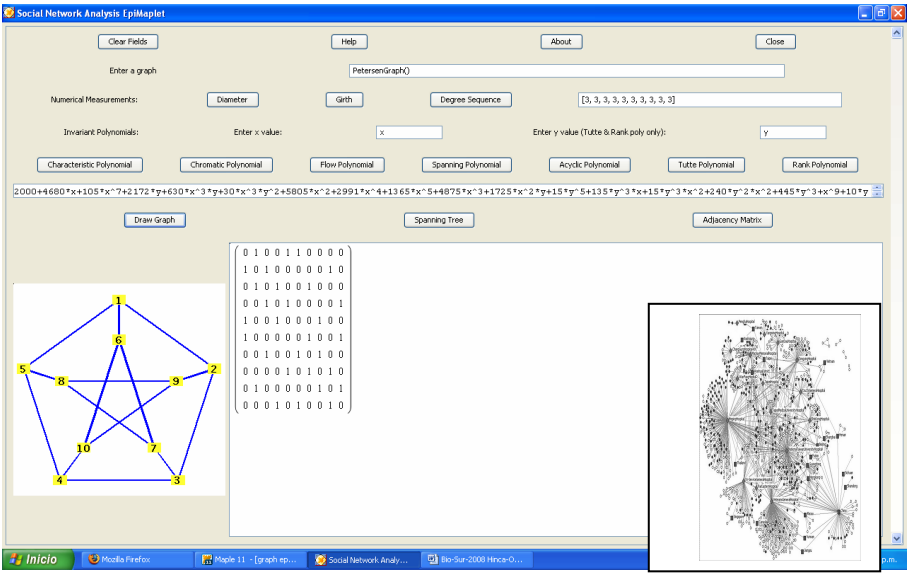


Fig. 7. Simulation of a Topological Quantum Computer using Computer Algebra

Actually, it is believed that quantum computers and specifically topological quantum computers are more efficient than classical computers when Tutte computations are involved. Recently, a topological quantum algorithm, named the Aharonov-Arab-Ebal-:Landau (AAEL algorithm) was proposed for the approximated computation of numerical evaluations of the Tutte polynomial for any given network [5,10]. As is

well known, the Tutte polynomial has an intrinsic relation to the Potts partition function and the AAEL algorithm is a step forward to the solution of the Potts problem, relevant in mathematical epidemiology and sociology.

A very rudimentary illustration of possible future applications of Topological Quantum Computers in Epidemiology, Bio-Surveillance and Bio-Security is depicted in Figure 6. This figure shows that a complex social network is considered as the input for a topological quantum computer that is able to compute the Tutte polynomial for the given network. Then the obtained Tutte polynomial is entered as the input for an intelligent standard quantum computer [11] that is able to make decisions in real time.

With the aim to have an idea of the possible future interface between epidemiologists and topological quantum computers, it is possible to build certain java applets that are designed using computer algebra, specifically maplets. Figure 7 shows a maplet which is able to compute the Tutte polynomial of a given social network with some complexity [12]. It is clear that the computer algebra is not able to replace Topological quantum computers, but the computer algebra is able to give some insight into the behavior of topological quantum computers.

4 Conclusions

In this work we have demonstrated that the Tutte polynomial codifies important combinatorial information of a given social network and that this combinatorial information is relevant in epidemiology. This information may indicate the way in which a social network is changing through time or give a characterization of the reduced networks resulting from other networks when some nodes are removed or show the effects on the basic reproductive number corresponding to spatial models with bi-networks, borders and memories. In this last case it is concluded that for the basic reproductive number, that incorporates the effects of borders, memories and bi-networks, the infectiousness is directly proportional to the Tutte polynomial of the contact network and it is possible to derive control measures to disrupt disease propagation from the borders to the interior of the country.

Consistent with this and according to the AAEL algorithm, topological quantum computers and algorithms may be a powerful tool for social network analysis applied in Epidemiology, Bio-Surveillance and Bio-Security.

A very interesting line for future research corresponds to the representation of the social networks as ribbon graphs and the subsequent application of the Bollobás-Riordan-Tutte polynomial for the epidemiological characterization of contact networks.

References

1. World Health Organization. The World Health Report 2007. A safer future: global public health security in the 21st century, Geneva (2007)
2. Eubank, S., Guclu, H., Kumar, A., et al.: Modelling disease outbreaks in realistic urban social networks. Nature 429, 180–184 (2004)

3. Chen, Y.-D., Tseng, C., King, C.-C., Wu, T.-S.J., Chen, H.: Incorporating Geographical Contacts into Social Network Analysis for Contact Tracing in Epidemiology: A Study on Taiwan SARS Data. In: Zeng, D., Gotham, I., Komatsu, K., Lynch, C., Thurmond, M., Madigan, D., Lober, B., Kvach, J., Chen, H. (eds.) BioSurveillance 2007. LNCS, vol. 4506, pp. 23–36. Springer, Heidelberg (2007)
4. Bollobás, B., Riordan, O.: A polynomial invariant of graphs on orientable surfaces. Proc. London Math. Soc. 83(3), 513–531 (2001)
5. Aharonov, D., Arad, I., Eban, E., Landau, Z.: Polynomial Quantum Algorithms for Additive aproximations of the Potts model and other Points of the Tutte Plane. In: QIP 2007, Australia (2007) arXiv:quant-ph/070208
6. Guyer, T.: Stratum and Compactness Maplet (December 2007), http://www.maplesoft.com/applications/app_center_view.aspx?AID=2195&CID=1&SCID=12
7. Champanerkar, A., Kofman, I., Stolzfus, N.: Quasi-tree expansion for the Bollobás-Riordan-Tutte polynomial (2007) arXiv:0705.3458v1
8. Hincapié, D., Ospina, J.: Spatial Epidemia Patterns Recognition using Computer Algebra. In: Zeng, D., Gotham, I., Komatsu, K., Lynch, C., Thurmond, M., Madigan, D., Lober, B., Kvach, J., Chen, H. (eds.) BioSurveillance 2007. LNCS, vol. 4506, pp. 216–221. Springer, Heidelberg (2007)
9. Jaeger, F., Vertigan, D.L., Welsh, D.J.A.: On the computational complexity of the Jones and Tutte polynomials. Mathematical Proceedings of the Cambridge Philosophical Society 108, 35–53 (1990)
10. Mario, V., Juan, O.: Possible quantum algorithms for the Bollobas- Riordan-Tutte polynomial of a ribbon graph. In: Donkor, E.J., Pirich, A.R., Brandt, H.E. (eds.) Proceedings of SPIE, March 27, 2008. Quantum Information and Computation VI, vol. 6976, pp. 697–60 (2008)
11. Tucci, R.R.: Use of a Quantum Computer and the Quick Medical Reference To Give an Approximate Diagnosis (2008) arXiv:0806.3949
12. Bodkin, P., Sherman, W.: Graph Analysis Maplet (November 2004), http://www.maplesoft.com/applications/app_center_view.aspx?AID=1394&CID=1&SCID=12

Integrating a Commuting Model with the Bayesian Aerosol Release Detector

Aurel Cami, Garrick L. Wallstrom, and William R. Hogan

Department of Biomedical Informatics, University of Pittsburgh,
Pittsburgh, PA, USA
{abc25,glw6,wrh9}@pitt.edu

Abstract. The Bayesian Aerosol Release Detector (BARD) is a biosurveillance system for detecting and characterizing disease outbreaks caused by aerosol releases of anthrax. A major challenge in modeling a population's exposure to aerosol anthrax is to accurately estimate the exposure level of each individual. In part, this challenge stems from the fact that the only spatial information routinely contained in the biosurveillance databases is the residential administrative unit (e.g., the home zip code of each case). To deal with this problem, nearly all anthrax biosurveillance systems, including BARD, assume that exposure to anthrax would occur at one's residential unit—a limiting assumption. We propose a refined version of BARD, called BARD-C, which incorporates the effect of commuting on a worker's exposure. We also present an experimental study to compare the performances of BARD and BARD-C on semi-synthetic outbreaks generated with an algorithm that also accounts for commuting.

Keywords: BARD, biosurveillance, anthrax, commuting.

1 Introduction

An outdoor aerosol release of *B. anthracis* could infect hundreds of thousands of individuals and without early detection mortality could be as high as 30,000 to 3 million [1, 2]. Therefore, the early detection of outbreaks caused by aerosol anthrax is an important problem in biosurveillance. The Bayesian Aerosol Release Detector (BARD) [3] is a system for detecting and characterizing such releases. BARD integrates the analysis of biosurveillance data—in the form of counts of Emergency Department (ED) visits with respiratory chief complaints (RCC)—with the analysis of meteorological and geographical data. Through this analysis BARD determines whether the current spatio-temporal pattern of respiratory disease incidence in the surveillance region is more consistent with the historical patterns or with the pattern that it would expect with an aerosol anthrax release.

Modeling a population's exposure to aerosol anthrax is a difficult problem. One major challenge is to estimate the exposure level (i.e., the number of inhaled anthrax spores) of each individual in the exposed region. The exposure level of an individual is largely determined by his location at the time of exposure. Hence, detailed spatial modeling at the person-level is required for accurate estimation of the exposure level. However, the data needed to parameterize such models are very difficult to obtain. In

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 85–96, 2008.

fact, the only type of spatial information that is routinely contained in biosurveillance databases is the residential administrative unit (e.g., the home zip code of each case). Faced with this lack of detailed spatial information, nearly all anthrax biosurveillance systems (including the existing version of BARD) make two common simplifying assumptions. First, they assume that all individuals that live in the same administrative unit would have the same exposure level—usually taken to be the level at the unit's centroid. Second, they assume that exposure to anthrax would occur at one's residential unit. While relaxing the first assumption remains currently an open challenge, in the last few years there has been some progress toward relaxing the second assumption. The key to this progress has been the integration of biosurveillance systems with mobility models that are parameterized through datasets that describe the travel patterns of the population. For one type of travel, namely the work-related commuting, such datasets are publicly available from the U.S. Census Bureau. For other types of travel, data are more difficult to obtain.

The first study that incorporated a mobility model in the simulation of anthrax outbreaks was conducted by Buckeridge [4]. He developed methods for integrating mobility in outbreak simulation and employed commuting data provided by the U.S. Census Bureau as well as survey-based non-commuting travel data. Then, he investigated the impact that incorporation of mobility in outbreak simulation had on two outbreak detection algorithms: a cumulative-sum temporal algorithm and the SMART spatial algorithm [5]. In a similar study, Cami et al. [6] investigated the impact that inclusion of commuting in outbreak simulation had on the detection and characterization performance of BARD. A few papers have investigated the integration of mobility models with outbreak detection algorithms. Duczmal and Buckeridge [7] proposed a method for integrating a commuting model with the spatial scan algorithm [8]. Garman et al. [9] developed a method for probabilistically inferring the work zip code from the home zip code and then integrated this method with the PANDA detection algorithm [10].

Here, we propose a simple and practical method for integrating a commuting model with BARD. We refer to the refined version of BARD that takes commuting into account as BARD-C. We present an experimental study that compares the performances of BARD and BARD-C on semi-synthetic outbreaks that were generated by a simulation algorithm that also employed a commuting model. We expected that BARD-C would perform better than BARD on these outbreaks. Our main research question was whether the improvement in performance would be large enough to warrant the additional computational cost.

2 An Overview of BARD

Next, we provide a brief description of BARD; an elaborate description of this system is given by Hogan et al. [3]. We focus on a region surrounding the city of Pittsburgh. This region consists of seven counties: Allegheny, Armstrong, Beaver, Butler, Lawrence, Washington, and Westmoreland. The ED-visit counts for the surveillance region are available at the granularity level of various administrative units, such as U.S. Census block groups and zip codes. Following the terminology in Lawson and Kleinman [11] we generically refer to these units as *tracts*.

2.1 Outbreak Detection with BARD

BARD attempts to discriminate between two hypotheses. The null hypothesis states that only "background" respiratory disease is present in the surveillance region. The alternative hypothesis states that both background respiratory disease and respiratory disease caused by an aerosol release of anthrax are present. Using Bayes' theorem, BARD computes the posterior probability of the attack hypothesis H_1 given data, as follows:

$$P(H_1 \mid \mathbf{b}, \mathbf{G}, \mathbf{M}) = \frac{P(\mathbf{b} \mid H_1, \mathbf{G}, \mathbf{M}) P(H_1)}{P(\mathbf{b} \mid H_0, \mathbf{G}, \mathbf{M}) P(H_0) + P(\mathbf{b} \mid H_1, \mathbf{G}, \mathbf{M}) P(H_1)}. \tag{1}$$

In addition, BARD computes the likelihood ratio, or the Bayes factor

$$\frac{P(\mathbf{b} \mid H_1, \mathbf{G}, \mathbf{M})}{P(\mathbf{b} \mid H_0, \mathbf{G}, \mathbf{M})}. \tag{2}$$

In equations (1)-(2), $\mathbf{b}$ is the *biosurveillance vector* consisting of the tract-level counts of ED visits with RCC during the last 24 hours; $\mathbf{G}$ is the *geographical matrix* containing the tract populations and the *x*, *y* coordinates of the tract centroids; $\mathbf{M}$ is the *meteorological matrix* containing the wind speed, the wind direction and the atmospheric class—a measure of atmospheric turbulence—for various observation times during the most recent week. The posterior probability of H_1 and the Bayes factor are both measures of the evidence provided by the data in favor of the alternative hypothesis.

The two key quantities in equations (1)-(2) are $P(\mathbf{b} \mid H_0, \mathbf{G}, \mathbf{M})$, the likelihood of biosurveillance data under H_0, and $P(\mathbf{b} \mid H_1, \mathbf{G}, \mathbf{M})$, the likelihood of biosurveillance data under H_1. Assuming conditional independence among the counts of different tracts, given a hypothesis and the meteorological conditions, BARD computes the two region-wide likelihoods $P(\mathbf{b} \mid H_0, \mathbf{G}, \mathbf{M})$ and $P(\mathbf{b} \mid H_1, \mathbf{G}, \mathbf{M})$ by first computing the corresponding tract-level likelihoods and then using equations:

$$\begin{aligned} P(\mathbf{b} \mid H_0, \mathbf{G}, \mathbf{M}) &= \prod_i P(c_i \mid H_0, \mathbf{G}, \mathbf{M}), \\ P(\mathbf{b} \mid H_1, \mathbf{G}, \mathbf{M}) &= \int_{\mathbf{r}} \left[\prod_i P(c_i \mid H_1, \mathbf{G}, \mathbf{M}, \mathbf{r}) \right] P(\mathbf{r} \mid H_1, \mathbf{G}, \mathbf{M}) \, d\mathbf{r}. \end{aligned} \tag{3}$$

In eq. (3), c_i denotes the 24-hr count of tract i and $\mathbf{r} = (x, y, h, q, t)$ denotes the release parameter vector, consisting of the release location x, y, h, the release quantity q, and the release time t. To compute the tract-level likelihoods $P(c_i \mid H_0, \mathbf{G}, \mathbf{M})$ and $P(c_i \mid H_1, \mathbf{G}, \mathbf{M}, \mathbf{r})$, BARD employs the binomial probability model.

The *null model* of each tract i is specified as follows:

$$\begin{aligned} c_i \mid \theta_{0,i} &\sim \mathrm{Bin}(n_i, \theta_{0,i}), \\ \theta_{0,i} &\sim \mathrm{Beta}(\alpha_i, \beta_i). \end{aligned} \tag{4}$$

In eq. (4), $\theta_{0,i}$ denotes the probability that a person who lives in tract i has visited an ED with RCC in the last 24 hours due to background disease; n_i denotes the population of tract i. By using the model specified by eq. (4) and integrating over the parameter $\theta_{0,i}$ it can be shown that the likelihood $P(c_i \mid H_0, \mathbf{G}, \mathbf{M})$ is a beta-binomial probability. BARD computes the estimates $\hat{\alpha}_i$, $\hat{\beta}_i$ of the beta distribution parameters from historical data for tract *i* through a moment-matching approach that takes into account the *day-of-week* and *month-of-year* variation of the baseline ED data.

The *alternative model* of each tract i is specified as follows:

$$\begin{aligned} c_i \mid \theta_{1,i} &\sim \mathrm{Bin}(n_i, \theta_{1,i}), \\ \theta_{0,i} &\sim \mathrm{Beta}(\alpha_i, \beta_i), \\ \theta_{1,i} &= 1 - (1 - \theta_{0,i})(1 - \theta_{1,i}^{+}). \end{aligned} \tag{5}$$

In eq. (5), $\theta_{1,i}$ denotes the probability that a person who lives in tract i has visited an ED with RCC in the last 24 hours either due to the background disease or due to exposure to anthrax; $\theta_{1,i}^{+}$ denotes the probability that a person who lives in tract i has visited an ED with RCC in the last 24 hours due to exposure to anthrax. The last identity in eq. (5) is derived by assuming causal independence between ED visits due to the background disease and ED visits due to exposure to anthrax.

BARD computes the quantity $P(c_i \mid H_1, \mathbf{G}, \mathbf{M}, \mathbf{r})$ by first deriving an estimate for the probability $\theta_{1,i}^{+}$ and then integrating over the parameter $\theta_{0,i}$. The estimate for $\theta_{1,i}^{+}$ is computed by taking into account (i) the *dose* of anthrax spores that would be observed in the centroid of tract *i* given the release scenario specified by $\mathbf{r}$, and (ii) a set of parameters for a *model of the respiratory disease* caused by inhalational anthrax. The dose of anthrax spores is computed through the Gaussian plume model of atmospheric dispersion (see [3]). The key disease-specific parameters that are taken into account in the estimation of $\theta_{1,i}^{+}$ are: (i) *minute ventilation*, the volume of air breathed per unit of time, (ii) the ID_{50}, the dose of spores infectious for 50% of the population, (iii) the *probit slope*, or the slope of the line that specifies the relationship between a probit—defined as Φ^{-1} (fraction of exposed who die), where Φ denotes the CDF of the standard normal distribution—and the logarithm of the dose of inhaled spores, and (iv) the parameters of a *log-normal distribution* that is employed to model the interval from exposure to the ED visit. Hogan et al. [3] give a detailed account of these parameters.

After computing $\theta_{1,i}^{+}$, BARD integrates with respect to $\theta_{0,i}$ to compute the quantity $P(c_i \mid H_1, \mathbf{G}, \mathbf{M}, \mathbf{r})$. Finally, BARD integrates over the release scenarios $\mathbf{r}$ through a Monte Carlo integration technique called *likelihood weighting* [3].

2.2 Outbreak Characterization with BARD

In addition to computing the posterior probability of H_1 and the Bayes factor, which can be used as alarm statistics for detecting outbreaks, BARD computes an estimate

for each element of the release vector $\mathbf{r}$. Accurate characterization of a release might assist responders in mitigating the impact of an outbreak. The estimation of $\mathbf{r}$ is carried out in Bayesian fashion. BARD assumes that the release parameters are conditionally independent given H_1, i.e.,

$$P(\mathbf{r} \mid H_1, \mathbf{G}, \mathbf{M}) = P(x, y \mid H_1)P(h \mid H_1)P(q \mid H_1)P(t \mid H_1)\,. \tag{6}$$

BARD employs a uniform prior for each element of $\mathbf{r}$, except h, for which a prior that favors smaller release heights relative to the higher ones is employed (see [3], p. 5248). Finally, BARD computes the posterior expectation of each element of $\mathbf{r}$ inside the *likelihood-weighting integration* procedure, mentioned in Section 2.2. The posterior expectation of $\mathbf{r}$ constitutes the release characterization produced by BARD.

As a final remark, BARD can also be used to *simulate* anthrax outbreaks. The BARD simulator [3] produces semi-synthetic outbreaks, created by first simulating cases due to an aerosol release of anthrax and then injecting the simulated cases into real ED-visit data. Cami et al. [6] developed a refined version of BARD simulator, called BARD-C simulator, which incorporates a commuting model in outbreak simulation.

3 Integration of a Commuting Model with BARD

The dataset that describes the commuting patterns is provided by the U.S. Census Bureau. This dataset was collected during the 2000 Census, has national coverage and is provided at the *census tract* level: each commuting flow represents the individuals who commute daily between a residence census tract and a work census tract. The commuting flows can be naturally modeled by a weighted, directed graph G in which nodes denote tracts, arcs represent commuting flows, and the weight of an arc denotes the number of commuters in the corresponding flow. We extracted from the nationwide commuting dataset the flows for which both the residence tract and the work tract belong to our surveillance region. The total number of commuters in this intra-region subset of flows was 1,005,566. Note that the flows for which only one of the two end-tracts belonged to the region accounted for only 3% of the intra-region flow and hence we ignored them for modeling convenience. In a pre-processing step we transformed the commuting flows from the census tract level provided by the Census Bureau to one of the two levels supported in BARD, namely the block group level. This pre-processing was carried out by splitting each commuting flow between a pair of census tracts T_1, T_2 into several smaller-sized flows, each joining a constituent block group of T_1 with a constituent block group of T_2, as discussed in [6]. The final commuting graph for our region consisted of 1991 nodes (block groups) and 324,402 arcs (commuting flows).

3.1 Development of BARD-C Detector

BARD-C detector takes as input a representation of the commuting graph G, in addition to the biosurveillance, meteorological, and geographical data. BARD-C computes

the likelihood $P(\mathbf{b} \mid H_0, \mathbf{G}, \mathbf{M})$ of biosurveillance data under the *null* hypothesis in exactly the same way as BARD. It could be argued that BARD's null model implicitly takes the commuting effect into account by parameterizing the beta distributions using historical data.

The refinement comes in the computation of the likelihood of biosurveillance data under the *alternative* hypothesis, i.e., $P(\mathbf{b} \mid H_1, \mathbf{G}, \mathbf{M})$. In developing the refinement needed to account for commuting, we were guided by two main objectives. The first objective was to adjust the computation of $P(\mathbf{b} \mid H_1, \mathbf{G}, \mathbf{M})$ so as to account for the commuting-induced non-uniformity in the exposure level of individuals living in the same tract. The second objective was to retain the practical value of BARD by ensuring that the running time of BARD-C on a commodity computer was shorter than a few hours (which is the typical frequency for running a detection algorithm in production).

BARD-C employs the same alternative model as BARD, but a refined method for computing the parameters of that model. In eq. (7), we have re-written the alternative model of a tract i, using the superscript * to highlight the difference in the computation of the model parameters between BARD and BARD-C.

$$\begin{aligned} c_i \mid \theta^*_{1,i} &\sim \mathrm{Bin}(n_i, \theta^*_{1,i}), \\ \theta_{0,i} &\sim \mathrm{Beta}(\alpha_i, \beta_i), \\ \theta^*_{1,i} &= 1 - (1 - \theta_{0,i})(1 - \theta^{+*}_{1,i}). \end{aligned} \tag{7}$$

The parameter $\theta^*_{1,i}$ still denotes the probability that a person who *lives* in tract i has visited an ED with RCC in the last 24 hours either due to the background disease or due to exposure to anthrax; the parameter $\theta^{+*}_{1,i}$ still denotes the probability that a person who *lives* in tract i has visited an ED with RCC in the last 24 hours due to exposure to anthrax. Because the commuting is now taken into account, the exposure tract of an individual could be different from the residence tract. Let us denote by $\theta_{1,i}$ the probability that a person who *is exposed* in tract i has visited an ED with RCC in the last 24 hours either due to the background disease or due to exposure to anthrax. Likewise, let us denote by $\theta^+_{1,i}$ the probability that a person who *is exposed* in tract i has visited an ED with RCC in the last 24 hours due to exposure to anthrax. Of course, if commuting is not taken into account, as in the existing version of BARD, we would have $\theta^*_{1,i} = \theta_{1,i}$ and $\theta^{+*}_{1,i} = \theta^+_{1,i}$, for all tracts i. Note that the four parameters $\theta_{1,i}, \theta^*_{1,i}, \theta^+_{1,i}, \theta^{+*}_{1,i}$ are *temporal* in the sense that each of them denotes the probability that an event has occurred in the last 24 hours. Our problem now has reduced to finding a method for computing the probabilities $\theta^{+*}_{1,i}$. Using a four-step approach discussed in the Appendix, we derived the following expression

$$\theta^{+*}_{1,i} = \sum_{j \in Out(i)} \frac{n_{ij}}{n_i} \theta^+_{1,j}. \tag{8}$$

In eq. (8), $Out(i)$ denotes the union of the set $\{i\}$ with the set of out-neighbors of tract i (i.e., the tracts where people who live in tract i work). Eq. (8) serves as the basis of the refinement performed in BARD-C. First, BARD-C computes the parameters $\theta_{1,i}^{+}$ for all tracts i using the Gaussian plume model and the disease-specific parameters (as explained earlier) and then computes the parameters $\theta_{1,i}^{+*}$ using eq. (8). BARD-C then uses the adjusted parameter $\theta_{1,i}^{+*}$ instead of the parameter $\theta_{1,i}^{+}$ to compute the integral $P(c_i \mid H_1, \mathbf{G}, \mathbf{M}, \mathbf{r})$.

This refinement is performed in the innermost loop of the likelihood weighting procedure. Since the computation of eq. (8) requires on average $\overline{d}$ multiplications and additions—where $\overline{d}$ denotes the average out-degree of the commuting graph G—it follows that the running time of BARD-C is given by

$$T_{BARD-C} = O(\overline{d} \times T_{BARD}). \tag{9}$$

For the block group-level commuting graph of the Pittsburgh region that we created from the census data, $\overline{d}$ is approximately 150. Considering that BARD takes approximately 10 minutes to run for the Pittsburgh region in a commodity single-processor computer, the running time of the just-described version of BARD-C (which is nearly 25 hours) is too long for practical purposes.

We found a way to reduce the running time of BARD-C to less than 1 hour without significantly reducing the accuracy of the computation. The first improvement in the time complexity of BARD-C was achieved by investigating the structure of the arc weights in the graph G. We noticed that, on average, for each tract i nearly 70% of the total out-going commuting flow was concentrated in the largest 1/3 of the outgoing flows. Hence, a reasonable method to reduce the average out-degree of G would be to sort the outgoing arcs of each tract in decreasing order and then take into account only the largest K flows, where K is a cutoff value. We used a cutoff value of 50 to obtain a reduction by a factor of 3 in the running time of BARD-C. The truncated graph still contained nearly 70% of the total flow in the original graph. To obtain an additional reduction in BARD-C's running time, we reduced the number of Monte Carlo repetitions in the likelihood weighting procedure from 200,000, which is the default used in BARD, to 20,000. With $\overline{d} = 50$ and 20,000 Monte Carlo repetitions, BARD-C takes nearly 45 minutes to run for the Pittsburgh region in a single-processor machine with a 3GHz processor and 2Gb of main memory.

4 Experimental Comparison of BARD and BARD-C

We performed an experiment to compare the performances of BARD and BARD-C on semi-synthetic outbreaks generated with the BARD-C simulator. The historical ED data for our experiment were provided by 10 EDs operated by one health system. The ED data represented nearly 30% of all ED visits in the surveillance region. We divided the total period spanned by the historical ED data into a training period (1 January 1999 to 31 December 2004), which was used to train BARD's detection algorithm, and a test period (1 January 2005 to 31 December 2005) used in the

evaluation experiment. Finally, the weather data was provided from the National Weather Service, while the populations and central zip codes came from the ESRI ArcGIS Desktop product.

4.1 Experimental Design

A total of 100 semi-synthetic outbreaks were generated using the BARD-C simulator. For all simulations, the quantity q was fixed at 0.5 kg. The release time t was chosen uniformly at random from the test period, i.e., the year 2005. The parameters x, y, and h were drawn from their prior distributions. Each semi-synthetic outbreak was supplied as input to the BARD and BARD-C detectors. The detectors were executed 42 times on each simulation in increments of 4 hours: the first execution began 4 hours after the release, the second execution 8 hours after the release, and so on.

The *detection performance* of both detectors was measured through the time-to-detection, or *timeliness,* defined as the interval from the release time to the detection time. The timeliness is typically measured as a function of the *alarm threshold*, i.e., the threshold of the alarm statistic employed by the detector to discriminate between the non-outbreak and outbreak situations. For a given threshold, an alarm is considered to be false if the alarm statistic exceeds the threshold prior to outbreak onset. The *false alarm rate* (for a given alarm threshold) is the number of false alarms that occur per unit of time. Plotting timeliness against the false alarm rate is known as Activity Monitoring Operating Characteristic (AMOC) analysis [12]. The *characterization performance* of both detectors was measured through t, x, y, h and q (*absolute*) *errors*. The t error is defined as the interval between the release time and the estimate of the release time produced by BARD. The x, y, h, and q errors are defined analogously. We plotted each characterization metric against the time interval from the release to the beginning of BARD's execution, which we call the *time to execution*.

Since BARD leverages a model of the respiratory disease caused by inhalational anthrax in both simulation and detection, a rigorous evaluation of BARD's performance requires a sensitivity analysis on each parameter of the disease model (these parameters were listed in Section 2.1). Such a sensitivity analysis was carried out by Hogan et al. [3]. The goal of our experiment, however, is to simply compare BARD with BARD-C, which differs from BARD only in that it adjusts the detection algorithm so as to account for commuting. Since a priori we do not expect commuting to interact with any of the disease-specific parameters, we believe that the difference between the performances of BARD and BARD-C would be the same regardless of the values of the disease-specific parameters used in simulation. For this reason, here we do not perform a sensitivity analysis with respect to the parameters of the disease model. Instead, in every simulation we set the disease-specific parameters at their *baseline* values, i.e., the values used in detection.

4.2 Results

Figure 1 shows the AMOC analysis for timeliness. As seen, BARD-C's timeliness is nearly two hours smaller than BARD's timeliness at every false alarm rate. A statistical test showed that, for each false alarm rate, the mean timeliness of BARD-C was statistically different from the mean timeliness of BARD, at the 0.05 level.

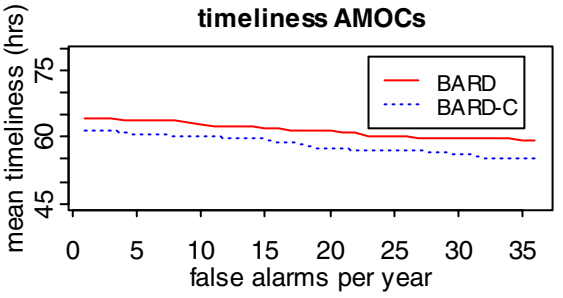


Fig. 1. AMOC curves for the timeliness of BARD and BARD-C detectors

Next, we turn to the characterization performance of BARD and BARD-C. First, we note that the performances of both detectors with respect to the q error and h error differed markedly from their performances with respect to the three remaining characterization metrics: $x, y,$ and t error. Unexpectedly, the posterior means of the release parameters q and h did not appear to converge as the time to execution increased. We leave the investigation of this unexpected outcome as a topic for future research. In the remainder of this section we focus on the x, y, and t errors.

Figure 2 shows box plots and plots of the medians of the x error for BARD and BARD-C. In each plot the horizontal axis shows the time-to-execution. The samples from which the box plots were constructed correspond to the different simulations. Several conclusions can be derived from Figure 2. First, the x error of both BARD and BARD-C converges to relatively small values as the time to execution increases. Second, the errors of BARD and BARD-C appear to converge around the same time and nearly 10-16 hours after the detection. Third, the sampling distribution of the x error is right-skewed for both detectors and in almost each execution of BARD and BARD-C there is a small number of outliers. Fourth, after the convergence, the median of the x error for BARD-C is nearly 2 kms smaller than the median of the x error for BARD.

Similar comments can be made for the y and t errors, whose plots are omitted due to space restriction. The post-convergence median of the y error for BARD-C was nearly 0.5 km smaller than the median for BARD; for the t error the post-convergence difference of the medians was nearly 20 minutes.

Figure 3 show the results of testing for sameness of characterization performance between BARD and BARD-C. The null hypothesis of this test asserted that BARD and BARD-C have identical locations of the distributions of x error (a), y error (b), and t error (c). The testing was carried out after each of the 42 runs of BARD and BARD-C. Due to the non-normality of the characterization errors, the Wilcoxon signed rank test was employed to perform the testing. Figure 3(a) plots the p-values corresponding to the x error against the time-to-execution. It can be seen that after the convergence of the x error, the p-values of the test remain well below the 0.05 significance level (the dotted horizontal line). Similar comments can be made for the y error. For the t error the convergence of the p values is not as clear as for the x and y errors.

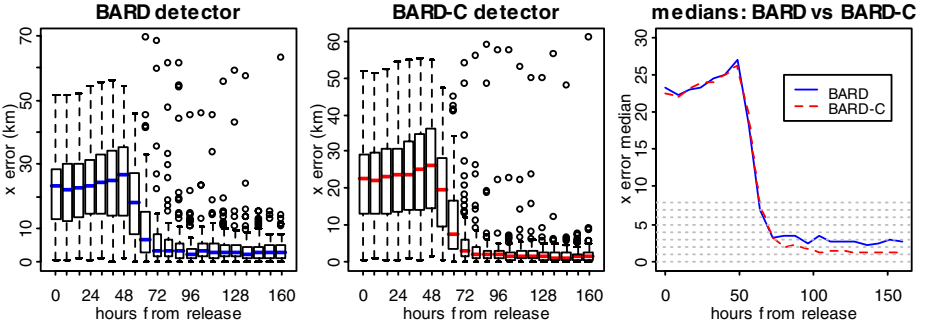
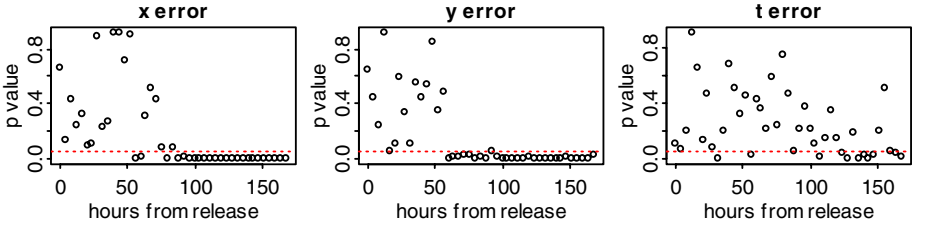


Fig. 2. Box plots and plots of the medians of the x error for BARD and BARD-C detectors



(a) (b) (c)

Fig. 3. P-values of the test for sameness of characterization performance between BARD and BARD-C as a function of the time-to-execution

5 Discussion

We proposed and evaluated a method for integrating a commuting model with BARD. To bound the running time of the refined detector, BARD-C, we made a number of simplifications at the expense of the accuracy of the computation. In spite of these simplifications, BARD-C performed better than BARD on a large set of semi-synthetic outbreaks that also incorporated commuting: BARD-C's timeliness was nearly 2 hours smaller than BARD's, BARD-C's x error was 1.5-2 kms smaller than BARD's and BARD-C's y error was nearly 0.5 km smaller than BARD's. In light of earlier studies [2], which estimated that a delay of just one hour in detection results in as much as $\$250$ million additional economic costs, this performance improvement is quite significant. We conclude that it is very important to study the problem further and, ultimately, to find the best tradeoff between efficiency and accuracy of the model.

Acknowledgments. We gratefully acknowledge the funding by the Centers for Disease Control and Prevention under grant R01PH000026-01.

References

[1] Kaufmann, A., Meltzer, M., Schmid, G.: The economic impact of a bioterrorist attack: are prevention and postattack intervention programs justifiable? Emerging Infectious Diseases 3, 83–94 (1997)

[2] Wagner, M.M., Tsui, F.C., Espino, J.U., Dato, V.M., Sittig, D.F., Caruana, R.A.: The emerging science of very early detection of disease outbreaks. Journal of Public Health Management Practice 7, 51–59 (2001)

[3] Hogan, W.R., Cooper, G.F., Wallstrom, G.L., Wagner, M.M., Depinay, J.-M.: The Bayesian aerosol release detector: An algorithm for detecting and characterizing outbreaks caused by an atmospheric release of Bacillus anthracis. Statistics in Medicine 26, 5225–5252 (2007)

[4] Buckeridge, D.L.: A method for evaluating outbreak detection in public health surveillance systems that use administrative data. PhD Thesis, Stanford University (2005)

[5] Kleinman, K.: Generalized linear models and generalized linear mixed models for small-area surveillance. In: Lawson, A.B., Kleinman, K. (eds.) Spatial and Syndromic Surveillance for Public Health, pp. 73–90. John Wiley & Sons, Chichester (2005)

[6] Cami, A., Wallstrom, G.L., Hogan, W.R.: Effect of commuting on the detection and characterization performance of the Bayesian aerosol release detector. In: Proc. International Workshop on Biomedical and Health Informatics, Philadelphia, PA, USA (November 2008) (to appear)

[7] Duczmal, L., Buckeridge, D.L.: A workflow spatial scan statistic. Statistics in Medicine 25, 743–754 (2006)

[8] Kulldorff, M.: A spatial scan statistic. Communications in Statistics: Theory and Methods 26, 1481–1496 (1997)

[9] Garman, C., Wong, W.K., Cooper, G.F.: The effect of inferring work location from home location in performing Bayesian biosurveillance. In: The Syndromic Surveillance Conference, Seattle, Washington (2005)

[10] Cooper, G.F., Dash, D.H., Levander, J.D., Wong, W.K., Hogan, W.R., Wagner, M.M.: Bayesian biosurveillance of disease outbreaks. In: Proc. Conference on Uncertainty in Artificial Intelligence, pp. 94–103 (2004)

[11] Lawson, A.B., Kleinman, K.: Spatial and Syndromic Surveillance for Public Health. Wiley, Chichester (2005)

[12] Fawcett, T., Provost, F.: Activity monitoring: noticing interesting changes in behavior. In: Proc. 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 1999), San Diego, California, United States, pp. 53–62. ACM Press, New York (1999)

Appendix

Here, we outline the steps we followed to derive eq. (8). In the subsequent derivation, in addition to the four temporal parameters $\theta_{1,i}, \theta^{*}_{1,i}, \theta^{+}_{1,i}, \theta^{+*}_{1,i}$ (Section 3.2) we will also need the two *non-temporal* parameters θ^{*}_{i} and θ_{i}. The former denotes the probability that a person who *lives* in tract i will *ever* visit an ED with RCC due to exposure to anthrax. The latter denotes the probability that a person who *is exposed* in tract i will *ever* visit an ED with RCC due to exposure to anthrax.

1. First, we derived a relationship between the *non-temporal* parameter θ^{*}_{i} and the *non-temporal* parameters θ_{j}, $j \in Out(i)$. To derive this relationship, given in eq. (10), we conditioned on the exposure tract.

$$\theta^{*}_{i} = \sum_{j \in Out(i)} \frac{n_{ij}}{n_{i}} \theta_{j} \,. \tag{10}$$

2. Second, it can be shown that the following relationship holds between the *temporal* parameter $\theta^{+}_{1,i}$ and the corresponding *non-temporal* parameter θ_{i}:

$$\theta^{+}_{1,i} = \theta_{i} \left[F(\tau \mid \mu_{i}, \sigma_{i}) - F(\tau - 1 \mid \mu_{i}, \sigma_{i}) \right]. \tag{11}$$

Here F denotes the CDF of a lognormal distribution with parameters μ_{i}, σ_{i} corresponding to the dose of spores in tract i, and τ is the interval that has elapsed from the hypothetical release time t to the beginning of the algorithm execution. Eq. (11) can be derived by conditioning (see [3], p. 5237).

3. Third, we derived a relationship between the *temporal* parameter $\theta^{+*}_{1,i}$ and the corresponding *non-temporal* parameter θ^{*}_{i}, using a combination of the conditioning techniques employed in the preceding two steps.

$$\theta^{+*}_{1,i} = \theta^{*}_{i} \left[\sum_{j \in Out(i)} \frac{n_{ij}}{n_{i}} \left(F(\tau \mid \mu_{j}, \sigma_{j}) - F(\tau - 1 \mid \mu_{j}, \sigma_{j}) \right) \right]. \tag{12}$$

4. Combining eqs. (10) and (12) yields

$$\theta^{+*}_{1,i} = \left[\sum_{j \in Out(i)} \frac{n_{ij}}{n_{i}} \theta_{j} \right] \left[\sum_{j \in Out(i)} \frac{n_{ij}}{n_{i}} \left(F(\tau \mid \mu_{j}, \sigma_{j}) - F(\tau - 1 \mid \mu_{j}, \sigma_{j}) \right) \right]. \tag{13}$$

Eq. (13) gives a rigorous method that can be employed in BARD-C to compute the adjusted parameters $\theta^{+*}_{1,i}$ in terms of quantities that are computed by BARD. In this paper, to reduce the time complexity of BARD-C, we simplified eq. (13) by assuming that $F(\tau \mid \mu_{j}, \sigma_{j}) - F(\tau - 1 \mid \mu_{j}, \sigma_{j})$ does not vary with j. This assumption leads to eq. (8) given in the body of the paper and to a reduction of the running time of BARD-C by at least a factor of 2.

A Temporal Extension of the Bayesian Aerosol Release Detector

Xiaohui Kong, Garrick L. Wallstrom, and William R. Hogan

Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260
{xik2,glw6,wrh9}@pitt.edu

Abstract. Early detection of bio-terrorist attacks is an important problem in public health surveillance. In this paper, we focus on the detection and characterization of outdoor aerosol releases of *Bacillus anthracis*. Recent research has shown promising results of early detection using Bayesian inference from syndromic data in conjunction with meteorological and geographical data [1]. Here we propose an extension of this algorithm that models multiple days of syndromic data to better exploit the temporal characteristics of anthrax outbreaks. Motivations, mechanism and evaluation of our proposed algorithm are described and discussed. An improvement is shown in timeliness of detection on simulated outdoor aerosol *Bacillus anthracis* releases.

Keywords: Anthrax outbreak, syndromic surveillance, Bayesian inference, spatial-temporal pattern recognition.

1 Introduction

In the event of an outdoor aerosol release of *B. anthracis*, it is critical that the release be detected and characterized quickly. A small amount of *B. anthracis* spores release could cause mortality in hundreds of thousands if not detected in a timely manner [2]. However, if it is detected early, vaccines and antibiotics could be deployed to significantly reduce mortality. A single hour of improvement in timeliness of detecting an aerosol release of *B. anthracis* could save as much as $250 million of economic cost [3]. In addition, early characterization of the release (location, time, affected area, etc.) enables public health intervention efforts to focus on the likely affected areas to further reduce mortality and economic cost. Syndromic surveillance is an alternative to case identification that uses less diagnostically-precise data from hospital emergency departments (EDs) to try to detect outbreaks earlier. *[5, 6]*. The common tools used by health departments to analyze ED visit data in order to detect disease outbreaks includes univariate time series analysis and spatial scan statistics [7-10]. However, these tools do not account for the unique pattern of disease that would likely result from an aerosol release of anthrax. In such a release scenario, weather conditions such as wind direction, wind speed, and atmospheric stability would influence the direction and shape of spore dispersal, and ultimately affect the location, shape and size of the affected area. The Bayesian Aerosol Release Detector (BARD) is an algorithm that uses meteorological data, in addition to syndromic and geographical data, to detect and characterize aerosol releases of anthrax [1]. Given any set of release parameters (time,

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 97–107, 2008.

location and quantity), BARD can calculate the probability of the observed ED visit counts over a spatial region in the last 24 hours using a dispersion model, an infection model and a visit delay model. BARD then calculates the posterior probability of a *B. anthracis* release by integrating all possible release scenarios with their corresponding prior probabilities and applying Bayes' rule. BARD also computes the expected release time, location and quantity by weighting each set of release parameters with their posterior probability.

One potential limitation of BARD is that only the ED visits in the most recent 24 were taken into account to detect outbreaks While historical visit data are utilized to specify prior distributions for background visit rates, recent visits that occurred more than 24 hours ago are not used to compute the posterior probability of a release or the expected release parameters. Hence BARD does not take full advantage of the temporal information available in the ED visit data.

In this paper we propose an extension of BARD called BARD-MD that models multiple days of data. As an extension of BARD, BARD-MD computes the posterior probability of an anthrax outbreak and its expected release parameters. We compare the performance of BARD-MD to BARD in terms of timeliness of detection, false alarm rate.

2 Methods

In this section we describe the model and inferential method of BARD-MD. We also indirectly describe BARD, as BARD is a special case of BARD-MD that uses n=1 days of ED visit data. Greater detail on the BARD model can be found in Hogan et al, 2007.

2.1 Data

BARD-MD uses three types of data for outbreak detection: biosurveillance data (B), geographic data (G), and meteorological data (M). The biosurveillance data are organized into an m x n matrix **B** containing the ED visit data for the most recent n days for the m zip codes in the surveillance data. Specifically, B_{ij} is the count of ED visits with respiratory complaints for zip code i during the time interval between 24×j hours ago and 24×(j-1) hours ago. The geographic data, represented by **G,** include the population and center coordinates for each zip code, and the mean and standard deviation of historical ED visit counts for each zip code, month of year and day of week combination in the training data. The meteorological data are arranged in a matrix **M** whose rows correspond to timestamps at which the meteorological observations were made and columns correspond to the particular variables observed. Specifically, the meteorological variables are wind speed, wind direction, and atmospheric stability class. Atmospheric stability class is a measurement of atmospheric turbulence, a key determinant of atmospheric dispersion of substances.

2.2 Hypotheses

BARD-MD entertains two mutually exclusive hypotheses. The null hypothesis H_0 is that 'background' respiratory disease that we have seen historically is the only 'respiratory disease' causing illness in the geographic region. The alternative

hypothesis H_1 is that both background respiratory disease and an outbreak of inhalational anthrax due to a release of *B. anthracis* are causing illness over the most recent n days. It is important to note that H_1 includes the event that the release occurred more than n days ago but caused illness within the most recent n days. The hypothesis H_1 also includes the event that the release occurred within the most recent n days even though background respiratory disease is alone responsible for respiratory visits prior to the release.

In addition to computing the posterior probability of H_1, BARD-MD computes the Bayes' Factor for H_1 compared to H_0. The Bayes' Factor is the ratio of the posterior odds of H_1 to the prior odds of H_1, and is equivalent to the marginal likelihood ratio of H_1 to H_0. (Kass & Raftery, 1995):

$$BF = \frac{P(H_1 \mid B,G,M) / P(H_0 \mid B,G,M)}{P(H_1 \mid G,M) / P(H_0 \mid G,M)} = \frac{P(B \mid H_1,G,M)}{P(B \mid H_{0,}G,M)} \tag{1}$$

The Bayes' Factor quantifies the evidence in favor of H_1 that was provided by the visit data, without requiring a careful assessment of the prior probabilities of H_0 and H_1. We now turn to the problem of calculating $P(B \mid H_{0,}G,M)$ and $P(B \mid H_1,G,M)$.

2.3 The Null Model

The model for B under the null hypothesis is described using an n x m matrix of parameters θ_0, where $\theta_{0,i,j}$ denotes the probability that a randomly selected resident of zip code i visited the ED with a respiratory complaint during the time interval between 24×j hours ago and 24×(j-1) hours ago. We assume that, under H_0, B is conditionally independent of M given G and θ_0: $P(B \mid H_0,G,M,\theta_0) = P(B \mid H_0,G,\theta_0)$. This assumption is tantamount to asserting that, in the absence of an aerosol release, if we know the historical mean and standard deviation of visit counts for the current month of year, day of week, and zip code, then the meteorological data provides no additional information about the background visit counts. We further assume that $\{b_{i,j}\}$ are conditionally independent under H_0 given the geographic data G and θ_0:

$$P(B \mid H_0,G,\theta_0) = \prod_{i=1}^{m}\prod_{j=1}^{n} P(b_{ij} \mid H_0,G,\theta_0) \tag{2}$$

Under H_0, each $b_{i,j}$ given G and θ_0 is modeled as a binomial process with size equal to the population of the zip code, n_i, and success probability $\theta_{0,i,j}$. We assume

that the $\theta_{0,i,j}$ values are conditionally independent given G and use a method of moments approach described in Hogan et al., 2007 to derive a beta prior distribution for each $\theta_{0,i,j}$ given G. See Hogan et al., 2007 for more details on prior specification. Observe that the assumption of conditional independence of $\theta_{0,i,j}$ is only reasonable when n <= 7, because the of the week effect is one of the factors taken into account to compute the beta priors. This assumption holds for the applications in this paper.

2.4 The Alternative Model

The alternative hypothesis H_1 is that there was an aerosol anthrax release that caused respiratory ED visits within the previous n days. We characterize the release with a parameter vector r that includes the time, location, and height of the release, as well as the quantity of anthrax spores released. The alternative model asserts that under $H_1, \{b_{ij}\}$ are conditionally independent given G, M, θ_0, and r, with each b_{ij} modeled as a binomial process. Specifically, the model asserts that

$$(b_{i,j} \mid H_1, G, M, \theta_0, r) \sim Bin(n_i, \theta_{1,i,j}) \tag{3}$$

where $\theta_{1,i,j}$ is a deterministic function of G, M, θ_0 and r that represents the probability under H_1 that an individual in zip code i visits ED with respiratory complaints j days ago. Because H_1 is the hypothesis that both background respiratory disease and an outbreak of inhalational anthrax are responsible for ED visits, we assume that these causes are independent. We compute $\theta_{1,i,j}$ as:

$$\theta_{1,i,j} = 1 - (1 - \theta_{0,i,j})(1 - \theta^{+}_{1,i,j}), \tag{4}$$

where $\theta^{+}_{1,i,j}$ is the probability that a resident of zip code i visits an ED with a respiratory complaint j days ago because of inhalational anthrax. When the release time is less than j days ago, $\theta^{+}_{1,i,j}$ is defined to be zero because the release cannot be responsible for ED visits prior to the release. Otherwise, when the release time is at least j days ago, $\theta^{+}_{1,i,j}$ is modeled as a function of the dose of spores *d* that an individual inhales and the amount of time *t* that has elapsed since he or she inhaled the spores. The dose *d* is derived from the Gaussian plume model of dispersion and an estimate of minute ventilation (the volume of air that an individual breathes per minute). Given any set of release parameters r, the Gaussian plume model of atmospheric dispersion uses meteorological and geographic data to compute the time-integrated concentration at an arbitrary downwind location due to a near-instantaneous release of a substance. The amount of time that elapses from spore inhalation to ED visit is specified by an infection model that relates the incubation period (time to symptom onset) to the dose of spores inhaled, and a visit delay model accounts for the amount of time that elapses from symptom onset to the ED visit. See

Hogan et al., 2007 for more detail on the Gaussian plume model, the infection model, and the visit delay model.

The prior distribution for θ_0 under H_1 is the same as under H_0. For the release parameters (release coordinates x, y, release height h, release amount Q and release time t), we assume that they are mutually independent under H_1 and independent of G, M and θ_0:

$$P(r \mid H_1, G, M, \theta_0) = P(x, y, \mid H_1) \square P(h \mid H_1) \square P(Q \mid H_1) \square P(t \mid H_1) \quad (5)$$

We use uniform priors for all of the release parameters except the height of the release, for which we use a probability function that decreases as values for h increase (see Appendix A3 in [1]). The location prior is uniform over the surveillance region, the quantity of released spores is uniform between 0.1 and 10 kilograms, and the time of release prior is uniform between 48 hours and 168 hours ago. As described in Hogan et al., the integral over θ_0 has a closed form solution (see Appendix A2 in [1]), and we use a Monte Carlo integration method called *likelihood weighting* to numerically integrate over **r** to calculate $P(B \mid H_1, G, M)$. The posterior expectation of the release location, height, time, and quantity are also calculated during the Monte Carlo integration.

3 Evaluation

We evaluate BARD-MD's ability to detect and characterize an aerosol anthrax release using semi-synthetic analysis in which anthrax releases are simulated and the resulting cases are added to real background ED visit data. We also compare the performance of BARD-MD when it runs every four hours using n=1 (BARD), n=2, and n=3 days of visit data using the same sets of semi-synthetic data.

3.1 Methods

Datasets. The evaluation region for this study is Pittsburgh, specifically a circular region centered on downtown Pittsburgh that contains 277 zip codes. The historical ED data used in this study were actual ED visits to ten EDs operated by one health system. We divided the historical ED visit data into training and test sets. The training set—which we used to parameterize BARD-MD's beta prior distributions—spanned a three-year time period from January 1, 2002 to December 31, 2004. The test set—into which we injected simulated anthrax-related ED visits—spanned one year from January 1, 2005 to December 31, 2005. The meteorological data we used are from the National Weather Service (NWS), which included wind speed and wind direction, but not stability class. We computed stability class with the available data using Turner's method [11]. The populations and central zip code points that we used in this study are from the ESRI® ArcGIS™ Desktop product.

Simulations. We simulated anthrax releases in the same way as BARD (Hogan et al., 2007). The following procedure was repeated fifty times to simulate fifty 0.1 kilogram releases and fifty 0.5 kilogram releases.

1. Select a release date and time uniformly from the interval 1/1/05-12/24/05, which is the interval covered by the test set of historical data but excluding the last seven days to ensure that the outbreak can fully manifest before the end of the test data. Also obtain the meteorological data for the selected date and time.

2. For release quantities of Q=0.1 kilograms and 0.5 kilograms, sample values for *x, y,* and *h* from their conditional prior distributions given Q and t. For the simulated outbreaks, we use the same non-uniform prior as BARD over *x* and *y*, which favors release locations desirable for their impact in terms of number of individuals infected. Specifically, the prior for the location of the release, conditional on the quantity, time and height of the release is proportional to the expected number of ED visits that would result from the release [1].

3. For each zip code in the region, use the Gaussian plume model, the infection model, and the visit delay model to simulate the number of ED anthrax visits and their presentation times. This simulation accounts for the facts that (1) not every case of inhalational anthrax would visit one of the 10 EDs in our historical dataset and (2) our historical dataset contains data on approximately 30% of ED visits in the Pittsburgh evaluation region.

Measurements. We measure false alarm rate, timeliness of detection and characterization of the releases for BARD-MD (two days and three days) and BARD. BARD-MD (two days) runs on the 24-hour aggregated ED counts from the last 48 hours. BARD-MD (three days) runs on the 24-hour aggregated ED counts from the last 72 hours. BARD runs on the aggregated ED counts of the last 24 hours. We defined a false alarm as an event when the Bayes' factor (likelihood ratio) of an outbreak exceeds an alarm threshold in the absence of an outbreak. The false alarm rate is then the number of false alarms that occur per year. To estimate the false-alarm rate, we ran BARD-MD (two days), BARD-MD (three days) and BARD with no simulated ED visits added to the baseline. These algorithms were run on the baseline data every four hours during January 1, 2005 to December 31, 2005. The algorithms were also run on the semi-synthetic data at the same four hour intervals a total of 42 times, starting from the first four-hour interval following the release and ending at a four-hour interval approximately seven days following the release. To measure timeliness of detection, we calculated the length of duration from simulated release to the first time when the Bayes' factor exceeded the alarm threshold. Then we took the mean of timeliness for each 50 simulations with 0.1 kg release and 0.5 kg release. For most event detection algorithms, there is a tradeoff on the performance between false alarm rate and timeliness of detection. Using the same method as previous researchers used to evaluate outbreak detection algorithms [1, 12, 13], we conducted an AMOC analysis [14], graphing the false-alarm rate versus time to detection. Our analysis used false alarm rates that ranged from 1 per year to 36 per year. Because we were interested in evaluating the effect of using multiple days' data vs. a single day's data for outbreak detection, we held all parameters in BARD constant across the simulations, BARD, and BARD-MD to isolate this effect. For a sensitivity analysis of BARD's performance, see [1].

4 Results

We compare the timeliness of detection and false alarm rate for BARD-MD (two days), BARD-MD (three days) and BARD in the AMOC plots below in Figure 1 for

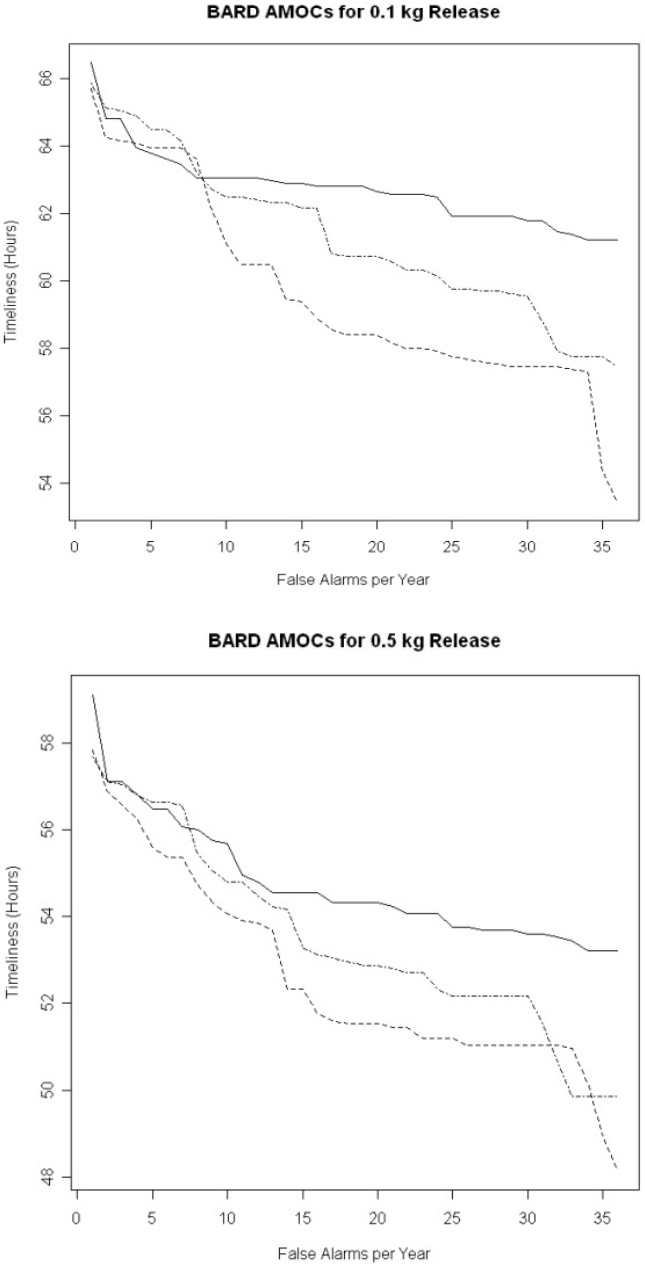


Fig. 1. Activity Monitoring Operator Characteristic (AMOC) curves for BARD (solid line), BARD-MD with two days of data (---), and BARD-MD with three days of data (.-.) for 0.1 kg releases (Figure 1a) and 0.5 kg releases (Figure 1b). The horizontal axis of each graph is the false alarm rate per year, and the vertical axis is the time to detection in hours.

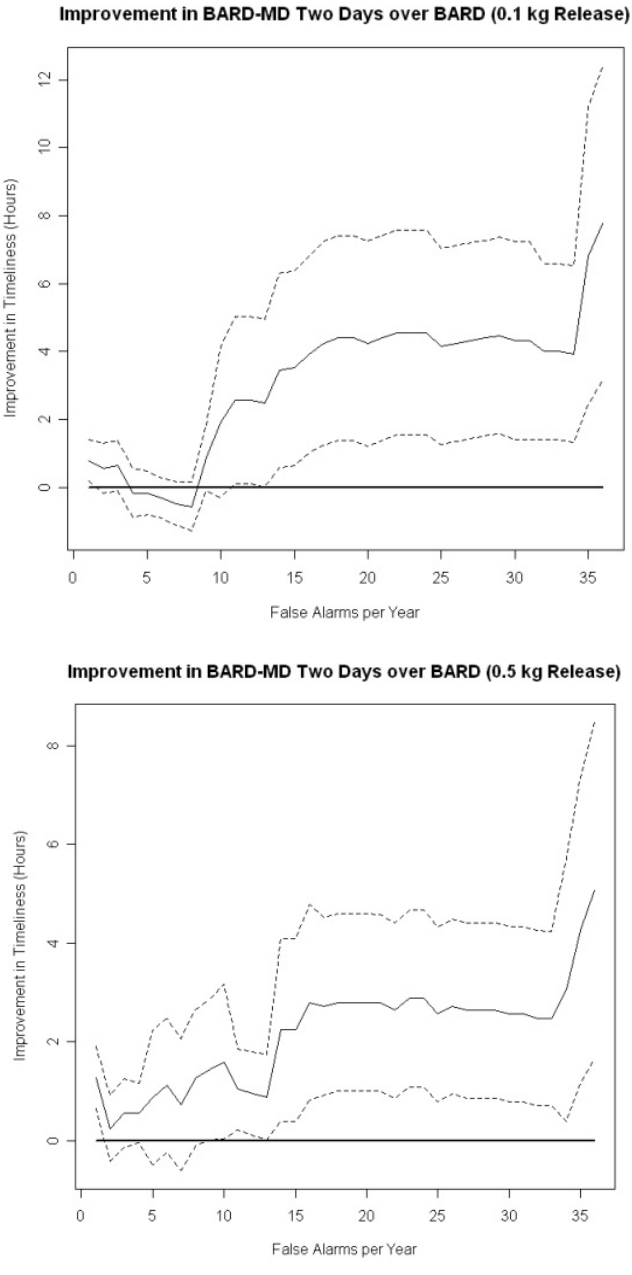


Fig. 2. 95% confidence bands for the difference in timeliness between BARD and BARD-MD with two days of data for 0.1 kg releases (Figure 2a) and 0.5 kg releases (Figure 2b)

0.1 kg and 0.5 kg releases. All algorithms successfully detected the 100 simulated outbreaks.

For both release quantities, our proposed algorithm BARD-MD using two days of data outperforms BARD in timeliness of detection when the number of false alarms is high and is about the same as BARD when the number of false alarms is low. At the high end when 36 false alarms are allowed, BARD-MD alarms 6 to 8 hours earlier than BARD. While BARD-MD with three days of data has better detection timeliness than BARD, it has worse timeliness than BARD-MD with two days of data. The remaining results that we present focus on the difference between BARD-MD with two days of data and BARD.

In Figure 2 we display point-wise confidence bands for the difference in timeliness between BARD and BARD-MD with two days of data. The confidence bands demonstrate that the improvement in timeliness is statistically significant at the 0.05 level at false alarm rates of 11 and higher per year for a 0.1 kg release, and at false alarm rates of 9 and higher per year for a 0.5 kg release.

5 Discussion

As we expected, taking more days of ED data into account does help in detecting simulated B. anthracis releases earlier. For lower false alarm rates, we did not find a statistically significant difference in the mean timeliness of detection for BARD-MD and BARD. It is unknown whether this finding is due to there being no real difference between the two algorithms at a lower false alarm rate, or due to using an insufficient number of simulation runs required to detect the difference. At higher false alarm rates, BARD-MD with two days of data detected simulated anthrax outbreaks earlier than BARD in our evaluation. We further investigated why BARD-MD with three days of data performs worse than BARD-MD with two days of data. The AMOC curves show (Figure 1), for BARD-MD the average timeliness of detection is 66 hours after release with the smallest false alarm rate of one per year. At that time of detection, BARD-MD with two days of data would use all ED visits that occur between 18 and 66 hours after the release while BARD-MD with three days of data would use visits that occur between -6 and 66 hours after the release. However, among the 3449619 simulated anthrax-related ED visits in the 100 outbreak scenarios *only one case* had an ED visit time within 24 hours of the release of *B. anthracis* spores. Thus, the additional day of data that is analyzed in the BARD-MD three days version compared to the BARD-MD two days version contains almost exclusively noise due to background respiratory disease and essentially no genuine signal of an anthrax outbreak.

Due to the rarity of real anthrax outbreak data, we evaluated both algorithms using synthetic data that was generated by injecting simulated anthrax-related ED visits into real baseline data. During a real anthrax outbreak, we would expect the detection time of both BARD and BARD-MD to be later than we found in our laboratory evaluation. Thus by the time of detection during a real outbreak, the ED visit data are likely to exhibit a stronger temporal pattern than in the laboratory evaluation, which we conjecture could result in a greater improvement in detection timeliness for BARD-MD compared to BARD.

6 Conclusion

To detect aerosol release of *B. anthracis* earlier, we developed an extension to the Bayesian Aerosol Release Detector algorithm, which we called BARD-MD. It uses multiple days of ED visit data to better account for temporal characteristics of an anthrax outbreak. We evaluated our algorithm's performance and compared it to BARD. We found that our proposed extension of BARD improves the timeliness of detection at high false alarm rates. Overall, BARD-MD is an important enhancement of BARD, and its use could result in reduced mortality and cost in the unfortunate event of an aerosol anthrax release.

Acknowledgements. This research was supported by a grant from the Centers for Disease Control and Prevention (R01PH000026). This work is solely the responsibility of its authors and do not necessarily represent the views of the CDC.

References

1. Hogan, W.R., et al.: The Bayesian aerosol release detector: an algorithm for detecting and characterizing outbreaks caused by an atmospheric release of Bacillus anthracis. Statistics in Medicine 26(29), 5225–5252 (2007)
2. Kaufmann, A.F., Meltzer, M.I., Schmid, G.P.: The economic impact of a bioterrorist attack: are prevention and postattack intervention programs justifiable? Emerging Infectious Diseases 3(2), 83–94 (1997)
3. Wagner, M.M., et al.: The emerging science of very early detection of disease outbreaks. Journal of Public Health Management & Practice 7(6), 51–59 (2001)
4. Buckeridge, D.L., et al.: Evaluating detection of an inhalational anthrax outbreak. Emerging Infectious Diseases 12(12), 1942–1949 (2006)
5. Heffernan, R., et al.: Syndromic surveillance in public health practice, New York City [erratum appears in Emerg Infect Dis. 2006 September 12(9):1472]. Emerging Infectious Diseases 10(5), 858–864 (2004)
6. Mandl, K.D., et al.: Implementing syndromic surveillance: a practical guide informed by the early experience. Journal of the American Medical Informatics Association 11(2), 141–150 (2004)
7. Lewis, M.D., et al.: Disease outbreak detection system using syndromic data in the greater Washington DC area [see comment]. American Journal of Preventive Medicine 23(3), 180–186 (2002)
8. Lombardo, J.: The ESSENCE disease surveillance test bed for the National Capital Area. Johns Hopkins APL Technical Digest, 327–334 (2003)
9. Tsui, F.C., et al.: Technical description of RODS: a real-time public health surveillance system. Journal of the American Medical Informatics Association 10(5), 399–408 (2003)
10. Wagner, M.M., et al.: Design of a national retail data monitor for public health surveillance. Journal of the American Medical Informatics Association 10(5), 409–418 (2003)
11. Tuner's Method (Accessed 2005 March 15) (2002), `http://www.webmet.com/met_monitoring/641.html`

12. Buckeridge, D.L., et al.: Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics 38(2), 99–113 (2005)
13. Wong, W.K., et al.: WSARE: What's Strange About Recent Events? Journal of Urban Health 80(2 suppl 1), 66–75 (2003)
14. Fawcett, T., Provost, F.: Activity monitoring: noticing interesting changes in behavior. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, San Diego (1999)

A Z-Score Based Multi-level Spatial Clustering Algorithm for the Detection of Disease Outbreaks

Jialan Que, Fu-Chiang Tsui, and Jeremy Espino

RODS Laboratory, Department of Biomedical Informatics, University of Pittsburgh
200 Meyran Street M-183, Pittsburgh, PA 15260
{jiq4,tsui2,juest4}@pitt.edu

Abstract. In this paper, we propose a Z-Score Based Multi-level Spatial Clustering (ZMSC) algorithm for the early detection of emerging disease outbreaks. Using semi-synthetic data for algorithm evaluation, we compared ZMSC with the *Wavelet Anomaly Detector* [1], a temporal algorithm, and two spatial clustering algorithms: *Kulldorff's spatial scan statistic* [2] and *Bayesian spatial scan statistic* [3]. ROC curve analysis shows that ZMSC has better discriminatory ability than the three compared algorithms. ZMSC demonstrated significant computational efficiency—1000x times faster than both spatial algorithms. Finally, ZMSC had the highest cluster positive predictive values of all the algorithms. However, ZMSC showed a 0.5-1 day average delay in detection when the false alarm rate was lower than one false alarm for every five days. We conclude that the ZMSC algorithm improves current methods of spatial cluster detection by offering better discriminatory ability, faster performance and more exact cluster identification.

Keywords: Spatial clustering, outbreak detection, biosurveillance.

1 Introduction

Disease outbreaks, either naturally occurring or caused by bioterrorism attacks, can result in significant morbidity, mortality and economic loss. Most outbreaks start in a small area and then expand to a larger area composed of geographically contiguous regions. For example, the SARS outbreaks of 2003 started in a Hong Kong apartment building and then quickly spread to all of Hong Kong and several countries in Asia. SARS claimed loss of hundreds of lives and cost billions, which leads to the urgency to detect outbreaks when they are small. This urgency has led to a subfield of epidemiology that systematically studies methods of spatial and temporal-spatial outbreak detection.

The two main approaches for outbreak detection are temporal analysis and spatial-temporal analysis. Temporal analysis using time series algorithms is the most popular approach. Time series algorithms include control chart, moving average (e.g., Exponentially Weighted Moving Average [4]), cumulative sum (CuSUM) [5], regressions [6], the Bayesian change-point detector [7], and the Wavelet Anomaly Detector (WAD) [1]. In this paper we compare the performance of ZMSC with WAD.

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 108–118, 2008.

WAD uses wavelet transform, a non-parametric algorithm suitable for non-stationary time series, to capture the underlying time series trend. An example of time series analysis is shown in Fig. 1 where two years and eight months of anti-diarrhea medication sales are analyzed using WAD to determine the underlying time series trend.

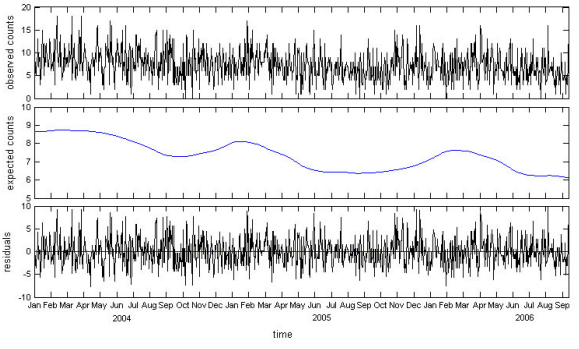


Fig. 1. A time series and its wavelet transform. The upper figure is a time series of over-the-counter (OTC) sales in the anti-diarrhea category over 2004-2006. The middle figure shows the expected values computed by wavelet transform. The bottom one is the residuals after subtracting the expected values from the observed values.

Spatial-temporal algorithms have been developed to take spatial distribution into account. In general the additional spatial distribution information allows spatial algorithms to achieve lower false alarm rates than temporal algorithms. Current state-of-the-art spatial algorithms include *Kulldorff's spatial scan statistic* (KSSS), *Bayesian spatial scan statistic* (BSSS) and Tango and Takahashi's *flexible spatial scan statistic* (FSSS). In addition, there are some other algorithms using classical clustering approaches, such as *Risk-adjusted Nearest Neighbor Hierarchical Clustering* (RNNH) [8] and *support vector machines* (SVMs) [9].

To focus on the new algorithm we developed, we briefly describe the three spatial-temporal algorithms based on scan statistics: KSSS, FSSS, and BSSS. KSSS [2, 9] scans a region by imposing circular or elliptic windows with different sizes, shapes and locations. The areas within a scanning window are considered a potential cluster. This algorithm finds a cluster with the highest likelihood ratio of having an outbreak in the cluster (H_1) vs. no outbreaks (H_0). FSSS [10, 11] is an improvement over KSSS by relaxing the artificial shape limitation of KSSS. It finds the cluster with any shape composed of k connected unit areas. BSSS [3] employs Bayes' rule to compute the posterior probability for each spatial region using a Poisson-Gamma model. The search window in BSSS is a rectangle (aligning with x and y axes) with varying width and height within a $m \times m$ grid. BSSS identifies a region with the greatest posterior probability of having an outbreak; thus, unlike KSSS, it does not require a randomization test.

There are some limitations to the KSSS, FSSS and BSSS. First, computation time for these algorithms is high. The three algorithms employ exhaustive searches for clusters, which dramatically increases computation time. Moreover, KSSS and FSSS require a randomization test to determine the significance of a detected cluster (i.e.,

p-value), which makes them intractable when the study region covers a large region such as one or more states. Secondly, KSSS and BSSS depend on the use of simple, fixed symmetrical shapes of regions. As a result, when the real underlying clusters do not conform to such shapes, the identified regions are often not well localized. [9]

Since it is public health interest to identify a clustered outbreak region based on temporal and spatial information, in this paper we will primarily focus on the performance comparison between ZMSC and two spatial-temporal algorithms--KSSS and BSSS--due to the wide acceptance of KSSS in public health surveillance and the recent innovative approach of BSSS. However, many biosurveillance systems still use time series algorithms for outbreak detection. Thus, we also include a time series algorithm WAD to serve as a baseline for the performance comparison in this study.

2 Methods

We propose a cluster detection algorithm that does not require exhaustive search. We first compute the z-score, to measure the degree of risk, for each area. We then find the subsets of the entire study area which have a high risk of outbreak. Finally, we identify clusters based on the adjacency relationship between any two areas in the subsets.

2.1 Z-Score Based Multi-level Spatial Clustering Algorithm (ZMSC)

Z-Score Based Risk Rate. We define Z_ρ as a subset of the entire study area ($\mathcal{Z}$). It is determined by a threshold value ρ as below:

$$Z_\rho = \{z_i : r_i \geq \rho, z_i \in \mathcal{Z}\} \tag{1}$$

where r_i represents the risk rate in area z_i. A risk rate can be computed as the ratio $r_i = c_i / b_i$ for area z_i, where c_i and b_i represent the observed cases and expected cases, respectively. However, such ratio only represents how far away the observed departs from the expected; it fails to take into account the degree of the deviation. Therefore, we propose in our model to compute the risk rate using z-score, $r_i = (c_i - b_i) / s_i$ instead, where s_i is an estimate of standard deviation σ of the residuals computed by subtracting the expected values from the corresponding observed values in time series. The computation of the risk rate or each unit study area is the first-stage analysis of ZMSC.

Subsets with Multiple Risk Levels. The areas in each sub-dataset are determined depending on the threshold value ρ for each risk level. ρ_p is assigned with the *p*-th percentile of the values in the set of $\{r_i, z_i \in \mathcal{Z}\}$, each of which represents the risk degree of having an outbreak in an area z_i. ZMSC sets $p = p_{\min}, ..., p_{\max}$ with intervals Δp. For instance, if $p_{\max} = 95\%$, $p_{\min} = 50\%$ and $\Delta p = 5\%$, the algorithm will get 10 sub-datasets with different risk levels ranging from 50-th percentile to 95-th percentile. This scheme only selects areas with elevated risks ($r_i \geq \rho_p$) to speed up cluster search.

Clustering on Subsets. For each subset Z_ρ, we divide individual areas into several clusters based on area adjacency constrain. We define an adjacency threshold η. The two areas z_i and z_j are said to be adjacent to each other (there is an edge between them) only if the minimal distance d_{ij} between these two areas is less than or equal to η (i.e. $d_{ij} \le \eta$). We restricted adjacency to connected areas in this paper (i.e., $\eta = 0$). Thus, if z_i and z_j share a borderline, the two areas become adjacent. The final cluster $\{z_i\}$, then, comprises a number of areas where for any two areas within the same cluster there is a path between them. For any two different clusters, there is no path from any area in one cluster to any area in the other.

Significance Analysis of a Cluster. The algorithm calculates the *p*-value of each cluster by combining all the normalized time series of its inclusive areas (e.g., ZIP codes). We test the hypothesis that there is an active elevated cluster happening on the current day within a particular region $S = \{z_i\}$ against there being no such cluster.

For each output cluster S we calculate the significance score p_S. We compose a new time series for S by summing up all the normalized time series in S. Each value in a normalized time series is a z-score, computed as $(c_{i,t} - \bar{c}_i)/s_i$, $t = 1,...,T$, where $c_{i,t}$ is the observed count in area z_i on day t; $\bar{c}_i = \sum_{t=1}^{T} c_{i,t}/T$ is the mean value of the time series of area z_i, and s_i is the estimated standard deviation of the time series. Hence, the new time series of region S with length T can be written as below:

$$F_S = (C_{S,1}, C_{S,2}, ..., C_{S,T}) = (\sum_{z_i \in S} \frac{c_{i,1} - \bar{c}_i}{s_i}, \sum_{z_i \in S} \frac{c_{i,2} - \bar{c}_i}{s_i}, ..., \sum_{z_i \in S} \frac{c_{i,T} - \bar{c}_i}{s_i}) \quad (2)$$

The algorithm applies wavelet transform to the combined normalized time series of each cluster to compute the corresponding expected value $B_{S,t}$. With the expected values available, the algorithm then computes z-scores of all previous days and the current day in time series, $R_{S,t} = (C_{S,t} - B_{S,t})/s_S$. Similar with individual areas, $R_{S,t}$ represents the degree of the deviation of the observed value from the expected in the aggregated region S at day t. The score p_S is computed as the ratio, $p_S = (M+1)/T$, where M is the number of values greater than or equal to $R_{S,T}$ in the set $\{R_{S,t}, t = 1,...,T-1\}$. A significant p_S indicates an emerging cluster occurring in the region S.

2.2 Algorithm Evaluation

OTC Pharmaceutical Sales Data. The dataset we used in this study contains 44 months of over-the-counter (OTC) anti-diarrhea medication sales data collected by the National Retail Data Monitor at the University of Pittsburgh with purchase dates

between Jan. 1, 2004 and Aug. 31, 2007 for the state of Pennsylvania. After filtering out the ZIP code areas with average daily sales less than 5 medications (these ZIP code areas had a large number of days with no sales) we identified 182 (out of 485) ZIP codes in this study. The evaluation period (12 months) was from Sep. 1, 2006 to Aug. 31, 2007 and the rest period was used for training and computation of expected values.

Semi-Synthetic Outbreaks. We injected a set of artificial outbreaks into the previously described OTC dataset to generate semi-synthetic experimental data. We randomly chose a group of adjacent ZIP codes from the collection of study ZIP codes. We used K to represent the size (the number of ZIP codes) of an outbreak and arbitrarily determined the outbreak's duration (e.g. $T=10$). The time of outbreak onset was randomly chosen as well within the evaluation period. Formula (3) computes the increased counts in each ZIP code during the outbreak period.

$$O(T,\delta,z_i)=\begin{cases}\delta\cdot t\cdot\bar{c}_i & 1\le t\le\left\lfloor\frac{T}{2}\right\rfloor\\ \delta\cdot(T-t)\cdot\bar{c}_i & \left\lfloor\frac{T}{2}\right\rfloor+1\le t\le T\end{cases} \tag{3}$$

where δ is the slope of the counts injected into the outbreak representing the outbreak strength, t is the number of the days in the outbreak, and $\bar{c}_i$ is the mean value of the time series of ZIP code z_i. The shape of a 10-day outbreak simulation is illustrated in Fig. 2.

In this study, we generated 4 groups of datasets, each of which had different outbreak settings, (K,δ), where $\delta\in\{0.2,0.3\}$ and $K\in\{4,8\}$. Each group included 100 datasets with different outbreaks, and each outbreak lasted for T=10 days. The results for the scenarios where the strength of the outbreaks was greater than 0.3 are not given in this paper because ZMSC could easily detect such outbreaks on the first day.

Algorithm Configuration. We compared ZMSC with one advanced time series algorithm (WAD) [1, 13] and two spatial-temporal algorithms–KSSS and BSSS. WAD has been evaluated in different studies comparing with different time series algorithms. KSSS has been well accepted and applied in public health surveillance field and BSSS is a recent innovative spatial-temporal algorithm using Bayesian approach.

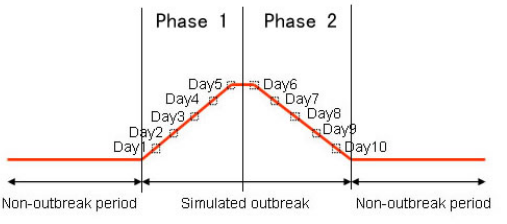


Fig. 2. An illustration of an artificial outbreak from Day 1 to Day 10

For KSSS, we applied the *space-time permutation model* [14] in *SaTScan*. We configured the analysis as *prospective*, and the time window as *1-day* (time precision was on a daily basis), which analyzes only the most current day to make it comparable to ZMSC. The number of Monte-Carlo replications was set to be *999*.

For BSSS, we calculated the expected values using a 28-day moving average (as described in [3]). The grid was defined as *32* by *32*.

For WAD, we applied wavelet transform to time series to calculate the expected value b_i for any area z_i and signaled an alarm if the value $(c_i - b_i)/s_i$ went beyond a pre-defined threshold on the most recent day T.

The parameters of the subsets to analyze in ZMSC were configured as $p_{\min} = 50\%$, $p_{\max} = 95\%$ and $\Delta p = 5\%$. The expected values b_i computed by wavelet transform were also used in ZMSC.

Experimental Environment. We ran all of our experiments on a 2GHz Intel CPU with 4G of memory. All of the algorithms were implemented in Java 1.5 except for KSSS. For KSSS we utilized the implementation in *SaTScan* which is coded in C.

Evaluation Metrics. We measured the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUROC), the activity monitoring operating characteristic (AMOC) curve, the area under the AMOC curve (AUAMOC), computation time, cluster sensitivity and cluster positive predictive value (PPV) for each algorithm and experimental dataset.

The AMOC curve represents the relationship between the false alarm rate (FAR) and the timeliness of outbreak detection. The FAR is the ratio of falsely detected outbreaks to all outbreaks signaled by the algorithm. The unit of timeliness measurement in this study was 1 day, assuming each algorithm is executed once a day. Both the area under the ROC and the area under the AMOC were ascertained using the trapezoidal approximation.

We defined cluster sensitivity as the percentage of the true outbreak areas over all outbreak areas. Each area in our study was the area within a ZIP code boundary. We defined cluster positive predictive value (PPV) as the ratio of the number of correctly detected outbreak areas in the cluster to the total number of areas in that cluster.

We defined a true positive as 1) at least one outbreak ZIP code is identified in the output cluster, and 2) the cluster is signaled within phase 1 of the outbreak (within the first 5 days of the outbreak).

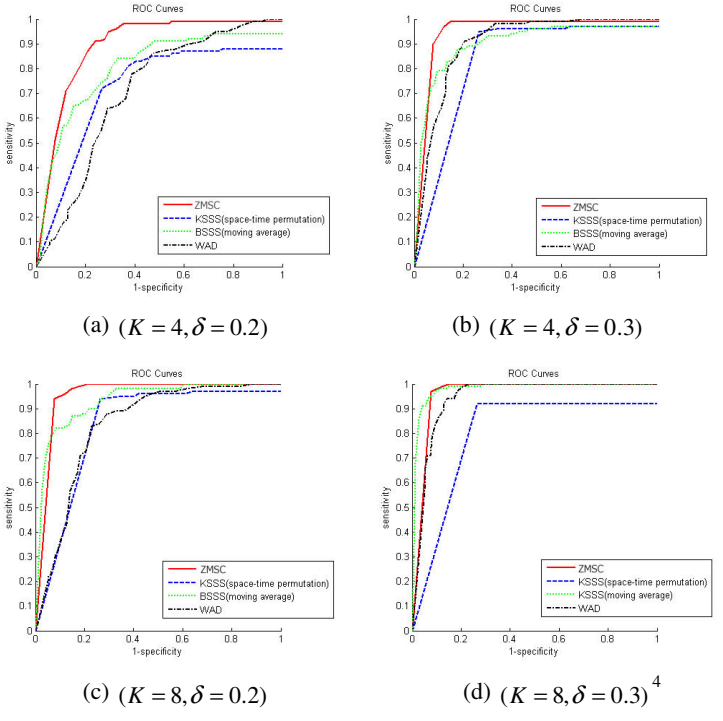
3 Results

Fig. 3 shows the ROC curves of the four algorithms (ZMSC, KSSS, BSSS and WAD) when run against data for four different types of synthetic outbreaks defined by different (K, δ). Table 2 shows the AUROC values for these experiments. ZMSC had the best (highest) area under the ROC (AUROC) in all experiments except for the group $(K = 8, \delta = 0.3)$. Fig. 4 shows the AMOC curves of the four algorithms. Table 2 shows the area under the AMOC values for these experiments. ZMSC had the best (lowest) AUAMOC in two groups of the experiments, ($K = 4, \delta = 0.2$) and

($K=8,\delta=0.2$). BSSS had the lowest AUAMOC in the remaining two groups of experiments, ($K=4,\delta=0.3$) and ($K=8,\delta=0.3$). We also found that KSSS in AMOC's consistently had a false alarm rate more than 0.2 per day.[1]

Table 1. Comparison of running time with 95% confidence intervals. Shaded cells show fastest algorithm based on 100 experiments.[2]

Algorithm	ZMSC	KSSS	BSSS	WAD
Running Time (seconds)[3]	0.186 (0.181-0.191)	933 (933-933)	1944 (1928-1959)	0.074 (0.074-0.074)



(a) ($K=4,\delta=0.2$) (b) ($K=4,\delta=0.3$)

(c) ($K=8,\delta=0.2$) (d) ($K=8,\delta=0.3$)[4]

Fig. 3. Comparison of ROC curves

[1] To compute the specificity of KSSS, we applied *SaTScan* to analyze the non-outbreak data on each day during a 1-year interval starting from Sep.1, 2006 to Aug. 31, 2007. KSSS output false clusters with p-values equal to 0.001 (999 randomization test) at more than 20% of the time.

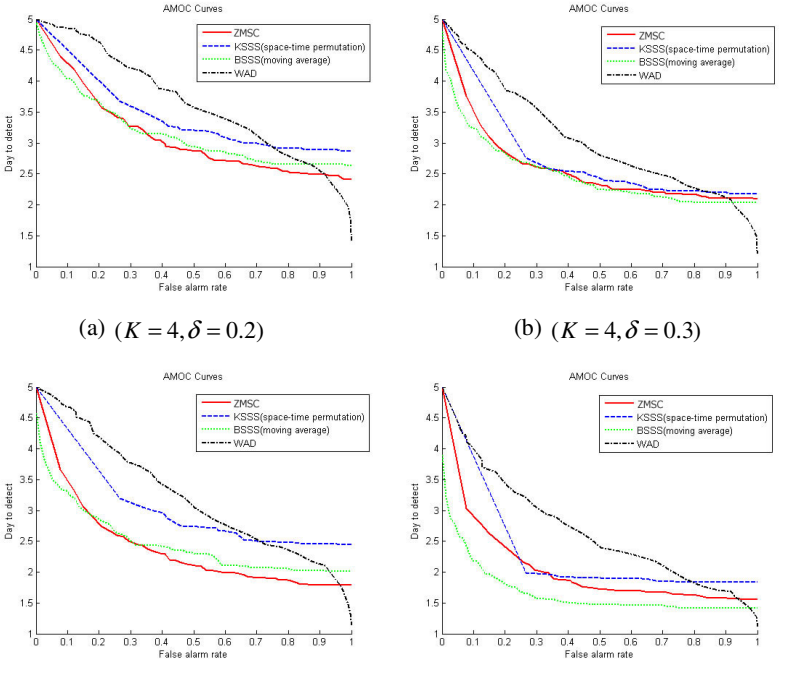
[2] The running times of KSSS (*SaTScan*) were quite close and did not differentiate too much which resulted in the identical upper control limit (ucl) and lower control limit (lcl) after rounding, as well as WAD.

[3] The KSSS method was executed using *SaTScan* (implemented in C), ZMSC, KSSS and BSSS were implemented in JAVA and executed under JRE-1.5. This comparison is meant to provide a rough idea of the running time of these algorithms.

[4] In this group of 100 experiments on KSSS, the sensitivity jumped to 0.92 when the specificity was lowered to 0.74 (i.e. 1-specificity=0.26) and kept unchanged.

Table 1 lists average running time of the three algorithms and the 95% confidence intervals. WAD ran the fastest. ZMSC ran 1000+ times faster than KSSS and BSSS.

Table 3 shows the average cluster sensitivities and PPV's (Positive Predictive Values) of the detected clusters. WAD has the highest cluster PPV overall. ZMSC has the highest PPV when compared to the other spatial algorithms (KSSS and BSSS). KSSS and BSSS had the highest cluster sensitivity values.



(a) $(K = 4, \delta = 0.2)$ (b) $(K = 4, \delta = 0.3)$

(c) $(K = 8, \delta = 0.2)$ (d) $(K = 8, \delta = 0.3)$

Fig. 4. Comparison of AMOC curves

Table 2. Comparison of the AUROC and AUAMOC. Shaded cells show the best performing algorithms for each group of experiments.

Algorithm		ZMSC	KSSS	BSSS	WAD
AUROC	$(K = 4, \delta = 0.2)$	0.89	0.72	0.81	0.71
	$(K = 4, \delta = 0.3)$	0.95	0.83	0.91	0.90
	$(K = 8, \delta = 0.2)$	0.96	0.83	0.94	0.84
	$(K = 8, \delta = 0.3)$	0.96	0.80	0.98	0.94
AUAMOC	$(K = 4, \delta = 0.2)$	3.11	3.44	3.15	3.65
	$(K = 4, \delta = 0.3)$	2.59	2.76	2.49	3.23
	$(K = 8, \delta = 0.2)$	2.39	3.05	2.46	3.06
	$(K = 8, \delta = 0.3)$	2.05	2.31	1.67	2.42

Table 3. Comparison of positive predictive values (PPV) and sensitivities of the clusters with 95% confidence intervals. Shaded cells show the best performing algorithms for each group of experiments.

	Cluster PPV				Cluster Sensitivity			
Algorithm	ZMSC	KSSS	BSSS	WAD	ZMSC	KSSS	BSSS	WAD
$(K = 4, \delta = 0.2)$	0.83 (0.78-0.87)	0.65 (0.65-0.66)	0.37 (0.30-0.44)	0.87 (0.78-0.96)	0.81 (0.77-0.85)	0.86 (0.86-0.87)	0.85 (0.80-0.90)	0.28 (0.25-0.32)
$(K = 4, \delta = 0.3)$	0.82 (0.77-0.86)	0.66 (0.65-0.67)	0.45 (0.38-0.51)	0.83 (0.78-0.89)	0.86 (0.82-0.89)	0.89 (0.88-0.89)	0.88 (0.83-0.92)	0.36 (0.32-0.39)
$(K = 8, \delta = 0.2)$	0.81 (0.77-0.85)	0.72 (0.71-0.72)	0.43 (0.38-0.48)	0.90 (0.85-0.94)	0.62 (0.57-0.67)	0.80 (0.80-0.80)	0.81 (0.76-0.85)	0.17 (0.15-0.19)
$(K = 8, \delta = 0.3)$	0.83 (0.79-0.88)	0.66 (0.65-0.67)	0.45 (0.40-0.50)	0.87 (0.81-0.92)	0.68 (0.64-0.73)	0.74 (0.74-0.75)	0.83 (0.79-0.86)	0.22 (0.19-0.25)

4 Discussion

In this study, WAD, a temporal algorithm, had the fastest running time and highest cluster PPV's among the four algorithms. The time complexity of wavelet pyramid algorithm is $O(n)$ which makes it the fastest algorithm within the four algorithms. Since it pinpoints the areas who have the highest elevated standard deviations, these areas were very likely to be the outbreak areas in the experiments (i.e. high cluster PPV's). However, the pinpointed areas by WAD were scattering because of the lack of spatial knowledge and the presence of noisy data, which leads WAD to the lowest cluster sensitivity compared to the other three algorithms which is not acceptable in practice for cluster detection. Hence, we pay more attention to the comparison among the three spatial detection algorithms in the following discussion. We found that ZMSC has several advantages—ZMSC is faster, more precise and able to detect arbitrarily shaped clusters.

First, ZMSC runs much faster than KSSS and BSSS due to its decreased computation complexity. Given the adjacency relationship between any two areas, the time complexity of the ZMSC algorithm is $O(n^2)$, where n is the number of ZIP codes in the analysis. ZMSC is more efficient than KSSS [2] which has a complexity of $O(n^3)$. ZMSC is also more efficient than BSSS [3] which has a complexity of $O(m^4)$, where m is the length of the grid.

Second, ZMSC tended to identify a cluster with much higher precision compared to the other algorithms. BSSS and KSSS had low PPV's, which indicates that their output clusters were either much larger than the sizes of true outbreaks or involved more non-outbreak areas than those of ZMSC. One of the reasons was that the shape-restricted search window (the cluster must be circular, ellipse or rectangular) involved more innocent areas. Another reason could be the inappropriate parameter setting of search, for example, 32 by 32 grid in BSSS for analyzing the state of Pennsylvania is not precise enough (but a denser grid setting makes it computational inefficient). ZMSC had the highest cluster PPV values while its sensitivity ranked third, which means that ZMSC would provide more focused areas for us to investigate further without eliminating too many outbreak areas.

Finally, ZMSC, unlike KSSS and BSSS, is not limited by the shape of a cluster. When ZMSC is used with ZIP code boundaries the cluster identified often has an irregular shape due to the union of the exact geographical shapes of the areas inside the output cluster. This approach is more informational than the others, which constrain the cluster shape into artificial shapes, such as circle, ellipse, or rectangle.

With regard to the area under the AMOC curves, the performance of ZMSC is comparable to the other three algorithms. However, ZMSC could not beat BSSS when the false alarm rate was low. Since the timeliness at low false alarm rates is more substantive in practice, this indicates BSSS had the best timeliness.

One limitation of this experiment is although most epidemic dispersions are more likely to propagate from starting areas to nearby (contiguous) areas, the ZMSC algorithm in this study was restricted to detecting outbreaks which were happening in the connected ZIP code areas only (adjacency threshold η was set to 0). For example, a cluster of outbreak areas separated by landforms such as rivers and mountains would probably be considered as several isolated smaller clusters by ZMSC. Such a limitation can be corrected by adjusting the adjacency threshold or taking landforms into account.

Also, as the injected outbreak was artificial in this study, it could not mimic perfectly real outbreaks. We plan to apply more challenging outbreak data in the future work.

Acknowledgments. This work is supported by grants NSF-IIS-0325581, CDC-1 R01 PH00026-01, NLM-5R21LM008278-03, PADOH-ME-01737, and AFRL-F30602-01-2-0550.

References

1. Zhang, J., Tsui, F.C., Wagner, M.M., Hogan, W.R.: Detection of Outbreaks from Time Series Data Using Wavelet Transform. In: AMIA Annual Symposium Proceeding, pp. 748–752 (2003)
2. Kulldorff, M.: A Spatial Scan Statistic. Communications in Statistics. Theory and Methods 26(6), 1481–1496 (1997)
3. Neill, D.B., Moore, A.W., Cooper, G.F.: A Bayesian Spatial Scan Statistic. In: Advances in Neural Information Processing Systems, vol. 18, pp. 1003–1010 (2005)
4. Hunter, J.S.: The Exponentially Weighted Moving Average. Journal of Quality Technology 18, 155–162 (1986)
5. Hawkins, D.M., Olwell, D.H.: Cumulative Sum Charts and Charting for Quality Improvement. Springer, Heidelberg (1998)
6. Wagner, M.M., Moore, A.W., Aryel, R.M.: Handbook of Biosurveillance. Elsevier, Amsterdam (2006)
7. Buckeridge, D.L., Burkom, H., Campbell, M., Hogan, W.R., Moore, A.W.: Algorithms for Rapid Outbreak Detection: a Research Synthesis. Journal of Biomedical Informatics 38(2), 99–113 (2005)
8. Jain, A., Murty, M., Flynn, P.: Data clustering: a review. ACM Computing Surveys 31(3), 264–323 (1999)

9. Kunihiko, T., Martin, K., Toshiro, T., Katherine, Y.: A Flexibly Shaped Space-Time Scan Statistic for Disease Outbreak Detection and Monitoring. International Journal of Health Geographics 7(14) (2008)
10. Kulldorff, M., Huang, L., Pickle, L., Duczmal, L.: An Elliptic Spatial Scan Statistic. Statistics in Medicine 25(22), 3929–3943 (2006)
11. Tango, T., Takahashi, K.: A Flexible Shaped Spatial scan Statistic for Detecting Clusters. International Journal of Health Geographics 4(11) (2005)
12. Zeng, D., Chang, W., Chen, H.: A Comparative Study of Spatio-Temporal Hotspot Analysis Techniques in Security Informatics. IEEE ITSC, 106–111 (2004)
13. Siegrist, D., Pavlin, J.: Bio-ALIRT biosurveillance detection algorithm evaluation. MMWR Morb Mortal Wkly Rep. 24(53) (suppl.), 152–158 (2004)
14. Kulldorff, M., Heffernan, R., Hartman, J., Assuncão, R., Mostashari, F.: A Space-Time Permutation Scan Statistic for Disease Outbreak Detection. PLoS medicine 2(3) (2005)

Epidemic Thresholds in SIR and SIIR Models Applying an Algorithmic Method

Doracelly Hincapié P.[1], Juan Ospina G.[2], Anthony Uyi Afuwape[3], and Ruben D. Gómez A.[1]

[1] Grupo de Epidemiología, Facultad Nacional de Salud Pública Universidad de Antioquia
doracely@guajiros.udea.edu.co, rdgomez@guajiros.udea.edu.co
[2] Grupo de Lógica y Computación, Universidad EAFIT
jospina@eafit.edu.co
[3] Grupo de Modelamiento en Ecuaciones Diferenciales, Universidad de Antioquia
aafuwape@matematicas.udea.edu.co

Abstract. Epidemic thresholds were deduced and simulated from SIR models of Susceptible – Infected – Recovered individuals, through local stability analysis of the disease free and endemic equilibrium, with an algorithmic method. One and two types of infected individuals were modeled, considering the influence of sub clinical, undiagnosed or unrecognized infected cases in disease transmission.

Keywords: Mathematical model, basic reproduction number.

1 Introduction

Recently, Brown et al. [1], proposed an algorithm for symbolic deduction of the basic reproductive rate through a local analysis of the disease-free state and endemic equilibrium.

The basic reproductive rate (R_0) is a critical magnitude or epidemic threshold that helps to understand the dynamics of emerging and re emerging disease transmission, identify measures to prevent and control epidemics and establish criteria for elimination / eradication of diseases. [2]

R_0 measures the average number of secondary cases generated by a primary case during its period of infectivity, when the case is introduced into a partially susceptible population. [3],[4]

When R_0 is the critical parameter deducted from a SIR model with homogeneous mixing between susceptible and infectious individuals, R_0 is a ratio between the infection rate of susceptible individuals and the recovered rate of infected individuals and multiplied by the susceptible size of population. [4]

Epidemic threshold is established according to R_0: If $R_0 > 1$, there will be instability of disease and outbreaks will occur because susceptible individuals accumulate long enough to start the outbreak or the infection rate is higher than the recovered rate. If $R_0 < 1$, there will be stability of disease, outbreak will be minor or will not occur at all because there are less susceptible individuals or there is a lower

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 119–130, 2008.

infection rate than the recovered rate by increasing of immunization, quarantine or mortality. [1]

A disease-free equilibrium is one in which all dependent variables corresponding to the presence of the disease in the population are zero. This equilibrium is asymptotically stable, if after a long period of time a state involving a small number of infected individuals will converge back to this disease-free equilibrium i.e., $R_0 <1$. It will be unstable, if secondary cases of the disease are generated i.e., $R_0 > 1$. [1]

Brown et al., analyzed the epidemic threshold in the following models in differential equations: the SEIRS model (susceptible, exposed –not yet infected-, infected, recovered –currently immuned-; the SEIT model adding a group T of individuals under treatment for the disease, the MSEIRS model whose newborn children of mothers (M) who are immune to a specific disease are passively protected by maternal antibodies for a certain time and the SIS model (susceptible, infected and vaccinated). [1]

These authors discussed the desirability of symbolic computation to analyze the properties of the parameters and influence of these in the epidemic threshold with an algorithmic approach that avoids tedious work by hand. [1]

This work continues Brown's algorithm by comparing the epidemic threshold in the SIR model with a single infected state and the SIIR model with two infected states. In both cases, the influence of immunization rate and loss of immunity rate are analyzed.

Modeling two infected states is important to understand the dynamics of transmission of sub clinical infections or asymptomatic cases, unrecognized or undiagnosed cases and diseases with different levels of severity. This is especially important when "undiagnosed" infected individuals may influence the transmission of infection, either by threatening the reemergence of the disease or limiting its elimination, such as influenza, SARS, polio, rubella, some sexually transmitted diseases, among others. [5], [6], [7]

2 Methods

An epidemic model is defined by a system of differential equations which describes the evolution of the number of individuals in each state of the epidemic process. [2], [8].

The SIR model reflects transitions from susceptible state to infected state when individuals have effective contact, according to the infection rate (β). Similarly, infected individuals are transferred to recovery state according to the recovery rate (γ), through isolation and recovery of infected individuals or through immunization of susceptible individuals [8].

In the SIIR model, susceptible individuals may be transferred to infected state number 1 (clinical, diagnosed, and recognized) or to the infected state number 2 (sub clinical, undiagnosed, and unrecognized), according to the infection rate β_1 and β_2, respectively. Similarly, infected individuals in each state are transferred to recovery state at the recovery rate γ_1 and γ_2. Immunization of susceptible individuals (p) and loss of immunity of recovered individuals (q) are analyzed in both models.

Throughout this paper we assume that birth and death rates (μ) are equal keeping a constant host population size, the population is homogeneously mixed and transmission is according to the mass-action principle. [8]

Epidemic thresholds are deducted through an analysis of local stability with a semiautomatic algorithm. [1] The algorithm is implemented in Maple 11 (Maplesoft Inc, Ontario Canada) and simulations are executed showing epidemic thresholds when there are changes of critical population size (N=10, N=100, N=1000). Packages for Groebner basis and Polynomial Ideals are exploited using as a background the power packages "LinearAlgebra" and "LargeExpressions".

2.1 The SIR Model

The system of equations describing the change in time of susceptible X(t), infected Y(t), and recovered Z(t) individuals, without immunization and loss of immunity rate, is:

$$\frac{d}{dt}\mathrm{X}(t) = \mu N - \beta\,\mathrm{X}(t)\,\mathrm{Y}(t) - \mu\,\mathrm{X}(t) \tag{1}$$

$$\frac{d}{dt}\mathrm{Y}(t) = \beta\,\mathrm{X}(t)\,\mathrm{Y}(t) - \gamma\,\mathrm{Y}(t) - \mu\,\mathrm{Y}(t) \tag{2}$$

$$\frac{d}{dt}\mathrm{Z}(t) = \gamma\,\mathrm{Y}(t) - \mu\,\mathrm{Z}(t) \tag{3}$$

2.2 The SIR Model with Immunization and Loss of Immunity

The differential equations are:

$$\frac{d}{dt}\mathrm{X}(t) = \mu N - \beta\,\mathrm{X}(t)\,\mathrm{Y}(t) - p\,\mathrm{X}(t) + q\,\mathrm{Z}(t) - \mu\,\mathrm{X}(t) \tag{4}$$

$$\frac{d}{dt}\mathrm{Y}(t) = \beta\,\mathrm{X}(t)\,\mathrm{Y}(t) - \gamma\,\mathrm{Y}(t) - \mu\,\mathrm{Y}(t) \tag{5}$$

$$\frac{d}{dt}\mathrm{Z}(t) = \gamma\,\mathrm{Y}(t) + p\,\mathrm{X}(t) - q\,\mathrm{Z}(t) - \mu\,\mathrm{Z}(t) \tag{6}$$

2.3 The SIIR Model

This model describes the epidemics with four states: Susceptible individuals X(t), Infected individuals of type 1- $\mathrm{Y}_1(t)$ (clinical, diagnosed, recognized), Infected individuals of type 2 – $\mathrm{Y}_2(t)$ (sub clinical, undiagnosed, unrecognized) and Recovered individuals Z(t). Immunization and loss of immunity rates are not included in this model.

$$\frac{d}{dt}\mathrm{X}(t) = \mu N - \beta_1 \mathrm{X}(t)\, Y_1(t) - \beta_2 \mathrm{X}(t)\, Y_2(t) - \mu \mathrm{X}(t) \tag{7}$$

$$\frac{d}{dt} Y_1(t) = \beta_1 \mathrm{X}(t)\, Y_1(t) - \gamma_1 Y_1(t) - \mu Y_1(t) \tag{8}$$

$$\frac{d}{dt} Y_2(t) = \beta_2 \mathrm{X}(t)\, Y_2(t) - \gamma_2 Y_2(t) - \mu Y_2(t) \tag{9}$$

$$\frac{d}{dt}\mathrm{Z}(t) = \gamma_1 Y_1(t) + \gamma_2 Y_2(t) - \mu \mathrm{Z}(t) \tag{10}$$

2.4 The SIIR Model with Immunization and Loss of Immunity

The corresponding system of equations is now:

$$\frac{d}{dt}\mathrm{X}(t) = \mu N - \beta_1 \mathrm{X}(t)\, Y_1(t) - \beta_2 \mathrm{X}(t)\, Y_2(t) - \mu \mathrm{X}(t) - p\, \mathrm{X}(t) + q\, \mathrm{Z}(t) \tag{11}$$

$$\frac{d}{dt} Y_1(t) = \beta_1 \mathrm{X}(t)\, Y_1(t) - \gamma_1 Y_1(t) - \mu Y_1(t) \tag{12}$$

$$\frac{d}{dt} Y_2(t) = \beta_2 \mathrm{X}(t)\, Y_2(t) - \gamma_2 Y_2(t) - \mu Y_2(t) \tag{13}$$

$$\frac{d}{dt}\mathrm{Z}(t) = \gamma_1 Y_1(t) + \gamma_2 Y_2(t) - \mu \mathrm{Z}(t) + p\, \mathrm{X}(t) - q\, \mathrm{Z}(t) \tag{14}$$

3 Results

3.1 Analysis of Local Stability

The points of diseases-free and endemic equilibrium of each model are presented in Table 1. The SIR models have a unique epidemic threshold with the presence of a single point of disease -free and endemic equilibrium, regardless of the presence of immunization and loss of immunity rates.

The SIIR models have both disease-free equilibrium states as endemic equilibrium states. This model exhibits two critical magnitudes corresponding to the basic reproductive rate of two sub-populations of infected individuals considered separately.

Details of the algorithm implementation are presented only to the SIIR model with immunization and loss of immunity rates.

Theorem 1. The system (11)-(14) admits the following equilibrium points:

a) $\{ Z = \frac{p\, N}{\mu + q + p},\ Y_1 = 0,\ Y_2 = 0,\ X = \frac{(\mu + q)\, N}{\mu + q + p} \}$

b)

$$\left\{ Y_2 = -\frac{\mu^2 + p\,\mu - \mu\,N\,\beta_2 + \mu\,\gamma_2 + \mu\,q - q\,N\,\beta_2 + q\,\gamma_2 + p\,\gamma_2}{\beta_2\,(\mu + \gamma_2 + q)},\ Y_1 = 0,\ X = \frac{\gamma_2 + \mu}{\beta_2},\right.$$
$$\left. Z = \frac{-\mu\,\gamma_2 + p\,\gamma_2 + p\,\mu + N\,\beta_2\,\gamma_2 - \gamma_2^{\,2}}{\beta_2\,(\mu + \gamma_2 + q)} \right\}$$

Table 1. Disease free and endemic equilibrium points and thresholds by SIR and SIIR model, with or without immuization rate (p) and loss of immunity rate (q)

Model	Disease free- equilibrium	Thresholds
SIR	$\{X = N, Z = 0, Y = 0\}$	$R_0 = \frac{N\beta}{\gamma + \mu}$
SIR pq	$Z = \frac{pN}{\mu + p + q}, Y = 0, X = \frac{(q + \mu)N}{\mu + p + q}$	$R_0 = \frac{N\beta\,(q + \mu)}{(\mu + p + q)\,(\gamma + \mu)}$
SIIR	$\{X = N, Y_2 = 0, Y_1 = 0, Z = 0\}$	$R_{0,1} = \frac{N\,\beta_1}{\gamma_1 + \mu}$ $R_{0,2} = \frac{N\,\beta_2}{\gamma_2 + \mu}$
SIIR pq	$Z = \frac{p\,N}{\mu + q + p}, X = \frac{(\mu + q)\,N}{\mu + q + p}, Y_1 = 0, Y_2 = 0$	$R_{0,1} = \frac{\beta_1\,(q + \mu)\,N}{(\mu + q + p)\,(\gamma_1 + \mu)}$ $R_{0,2} = \frac{\beta_2\,(q + \mu)\,N}{(\mu + q + p)\,(\gamma_2 + \mu)}$

Model	Endemic equilibrium	Thresholds
SIR	$Z = \frac{\gamma\,(N\beta - \gamma - \mu)}{\beta\,(\gamma + \mu)}, Y = \frac{\mu\,(N\beta - \gamma - \mu)}{\beta\,(\gamma + \mu)}, X = \frac{\gamma + \mu}{\beta}$	$R_0 = \frac{N\beta}{\gamma + \mu}$
SIR pq	$X = \frac{\gamma + \mu}{\beta}, Z = -\frac{-p\,\gamma - p\,\mu - N\,\beta\,\gamma + \gamma^2 + \mu\,\gamma}{\beta\,(\gamma + q + \mu)},$ $Y = -\frac{p\,\gamma + p\,\mu - q\,N\,\beta + q\,\gamma + \mu\,q - \mu\,N\,\beta + \mu\,\gamma + \mu^2}{(\gamma + q + \mu)\,\beta}$	$R_0 = \frac{N\beta\,(q + \mu)}{(\mu + p + q)\,(\gamma + \mu)}$
SIIR	a) $\{Y_1 = -\frac{\mu\,(-N\beta_1 + \gamma_1 + \mu)}{\beta_1\,(\gamma_1 + \mu)}, Z = -\frac{\gamma_1\,(-N\beta_1 + \gamma_1 + \mu)}{\beta_1\,(\gamma_1 + \mu)}, Y_2 = 0, X = \frac{\gamma_1 + \mu}{\beta_1}\}$ b) $\{X = \frac{\gamma_2 + \mu}{\beta_2}, Y_1 = 0, Y_2 = -\frac{\mu\,(-N\beta_2 + \gamma_2 + \mu)}{\beta_2\,(\gamma_2 + \mu)}, Z = -\frac{\gamma_2\,(-N\beta_2 + \gamma_2 + \mu)}{\beta_2\,(\gamma_2 + \mu)}\}$	a) $R_{0,1} := \frac{N\,\beta_1}{\gamma_1 + \mu}$, b) $R_{0,2} := \frac{N\,\beta_2}{\gamma_2 + \mu}$
SIIR	a)	a)

Table 1. (*continued*)

pq	$\left\{ Y_1 = -\frac{\mu^2 + p\,\mu - \mu\,N\,\beta_1 + \mu\,\gamma_1 + \mu\,q - q\,N\,\beta_1 + q\,\gamma_1 + p\,\gamma_1}{\beta_1\,(\mu + \gamma_1 + q)},\ Y_2 = 0,\ X = \frac{\gamma_1 + \mu}{\beta_1},\right.$ $\left. Z = \frac{-\mu\,\gamma_1 + p\,\gamma_1 + p\,\mu + N\,\beta_1\,\gamma_1 - \gamma_1^2}{\beta_1\,(\mu + \gamma_1 + q)} \right\}$	$R_{0,1} = \frac{\beta_1\,(\mu + q)\,N}{(\gamma_1 + \mu)\,(\mu + q + p)}$
b)	$\left\{ Y_2 = -\frac{\mu^2 + p\,\mu - \mu\,N\,\beta_2 + \mu\,\gamma_2 + \mu\,q - q\,N\,\beta_2 + q\,\gamma_2 + p\,\gamma_2}{\beta_2\,(\mu + \gamma_2 + q)},\ Y_1 = 0,\ X = \frac{\gamma_2 + \mu}{\beta_2},\right.$ $\left. Z = \frac{-\mu\,\gamma_2 + p\,\gamma_2 + p\,\mu + N\,\beta_2\,\gamma_2 - \gamma_2^2}{\beta_2\,(\mu + \gamma_2 + q)} \right\}$	b) $R_{0,2} := \frac{\beta_2\,(q + \mu)\,N}{(\mu + q + p)\,(\gamma_2 + \mu)}$

c)

$$\left\{ Y_1 = -\frac{\mu^2 + p\,\mu - \mu\,N\,\beta_1 + \mu\,\gamma_1 + \mu\,q - q\,N\,\beta_1 + q\,\gamma_1 + p\,\gamma_1}{\beta_1\,(\mu + \gamma_1 + q)},\ Y_2 = 0,\ X = \frac{\gamma_1 + \mu}{\beta_1},\right.$$
$$\left. Z = \frac{-\mu\,\gamma_1 + p\,\gamma_1 + p\,\mu + N\,\beta_1\,\gamma_1 - \gamma_1^2}{\beta_1\,(\mu + \gamma_1 + q)} \right\}$$

Proof:
Equations of equilibrium

$$\mu\,N - \beta_1\,X\,Y_1 - \beta_2\,X\,Y_2 - \mu\,X - p\,X + q\,Z = 0 \tag{15}$$

$$\beta_1\,X\,Y_1 - \gamma_1\,Y_1 - \mu\,Y_1 = 0 \tag{16}$$

$$\beta_2\,X\,Y_2 - \gamma_2\,Y_2 - \mu\,Y_2 = 0 \tag{17}$$

$$\gamma_1\,Y_1 + \gamma_2\,Y_2 - \mu\,Z + p\,X - q\,Z = 0 \tag{18}$$

Resolving (15)-(18) was obtained a), b) y c)

Theorem 2. In the system (11)-(14), the disease – free equilibrium point is locally stable if and only if, $R_{0,1}<1$ and $R_{0,2}<1$, where

$$R_{0,1} = \frac{\beta_1\,(q + \mu)\,N}{(\mu + q + p)\,(\gamma_1 + \mu)}, \qquad R_{0,2} = \frac{\beta_2\,(q + \mu)\,N}{(\mu + q + p)\,(\gamma_2 + \mu)}$$

Proof:
The Jacobian of the system (11)-(14) evaluated at the disease – free equilibrium point is:

$$\begin{bmatrix} -\mu - p & -\dfrac{\beta_1 (\mu + q) N}{\mu + q + p} & -\dfrac{\beta_2 (\mu + q) N}{\mu + q + p} & q \\ 0 & \dfrac{\beta_1 (\mu + q) N}{\mu + q + p} - \gamma_1 - \mu & 0 & 0 \\ 0 & 0 & \dfrac{\beta_2 (\mu + q) N}{\mu + q + p} - \gamma_2 - \mu & 0 \\ p & \gamma_1 & \gamma_2 & -\mu - q \end{bmatrix},$$

and the corresponding stability conditions are

$$0 < \mu^2 + p\,\mu - \mu\,N\,\beta_1 + \mu\,\gamma_1 + \mu\,q - q\,N\,\beta_1 + q\,\gamma_1 + p\,\gamma_1$$

$$0 < \mu^2 + p\,\mu - \mu\,N\,\beta_2 + \mu\,\gamma_2 + \mu\,q - q\,N\,\beta_2 + q\,\gamma_2 + p\,\gamma_2 .$$

These two stability conditions can be rewritten respectively as $R_{0,1} < 1$ and $R_{0,2} < 1$, where

$$R_{0,1} = \frac{\beta_1 (\mu + q) N}{(\gamma_1 + \mu)(\mu + q + p)}, \qquad R_{0,2} = \frac{\beta_2 (\mu + q) N}{(\gamma_2 + \mu)(\mu + q + p)} .$$

Theorem 3. In the system (11)-(14), the first point of the endemic equilibrium is locally stable when it exists, that is, when $R_{0,2}>1$, where

$$R_{0,2} := \frac{\beta_2 (q + \mu) N}{(\mu + q + p)(\gamma_2 + \mu)}$$

Proof:
The Jacobian for the first endemic equilibrium point is:

$$AAIIpq2 := \begin{bmatrix} \dfrac{\mu\,q - q\,N\,\beta_2 + q\,\gamma_2 + \mu^2 + p\,\mu - \mu\,N\,\beta_2 + \mu\,\gamma_2 + p\,\gamma_2}{\mu + q + \gamma_2} - \mu - p & -\dfrac{\beta_1 (\gamma_2 + \mu)}{\beta_2} & -\mu - \gamma_2 & q \\ 0 & \dfrac{\beta_1 (\gamma_2 + \mu)}{\beta_2} - \gamma_1 - \mu & 0 & 0 \\ -\dfrac{\mu\,q - q\,N\,\beta_2 + q\,\gamma_2 + \mu^2 + p\,\mu - \mu\,N\,\beta_2 + \mu\,\gamma_2 + p\,\gamma_2}{\mu + q + \gamma_2} & 0 & 0 & 0 \\ p & \gamma_1 & \gamma_2 & -q - \mu \end{bmatrix}$$

and the corresponding stability condition is: $R_{0,2} >1$; where

$$R_{0,2} := \frac{\beta_2 (q + \mu) N}{(\mu + q + p)(\gamma_2 + \mu)} .$$

Finally, given that

$$Y_2 = \frac{(\gamma_2 + \mu)(R_{0,2} - 1)(\mu + q + p)}{\beta_2 (\mu + \gamma_2 + q)},$$

The condition of existence of the first endemic state is $R_{0,2}>1$.

Theorem 4. In the system (11)-(14), the second point of the endemic equilibrium is locally stable when it exists, that is, when $R_{0,1}>1$, where

$$R_{0,1} = \frac{\beta_1 (\mu + q) N}{(\gamma_1 + \mu)(\mu + q + p)}$$

Proof: In analogy with the demonstration of Theorem 3.

3.2 Numerical Simulations

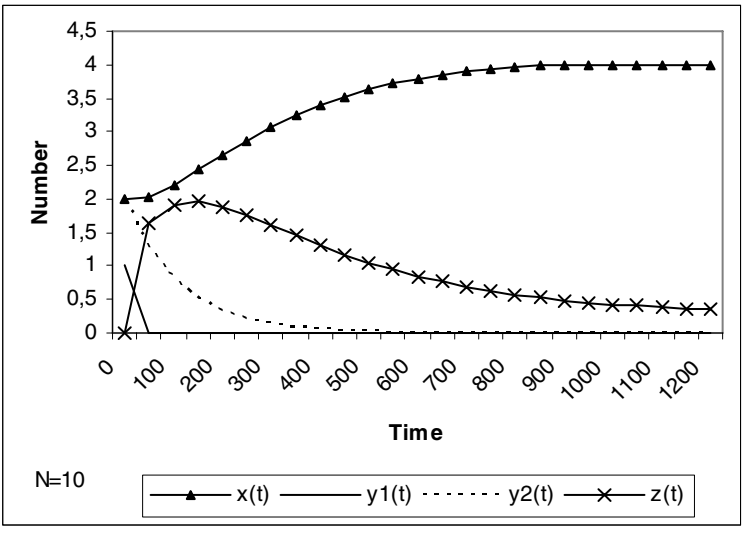
The Table 2 shows numerical simulation of epidemic thresholds and mathematical expresions for $y_1(t)$ clinical and $y_2(t)$ subclinical cases, with different critical population sizes and according to Theorem 2. Parameter values correspond to data from rubella (infection rate ~ incidence rate) in Latin America and the Caribbean in 1998, a few years after the start of mass vaccination against rubella. It is assumed a relationship 2:1 of clinical to subclinic infection, because 30-40% of rubella cases are subclinical.[9]

In the first simulation, $R_{0,1} < 1$ and $R_{0,2} <$, there is not epidemic outbreak. The Figure 1 a) shows the corresponding epidemic curves and the typical behaviour of stability are observed: the number of infected individuals is decreased to zero and finally only susceptible and recovered individuals remain; which means there is not an outbreak.

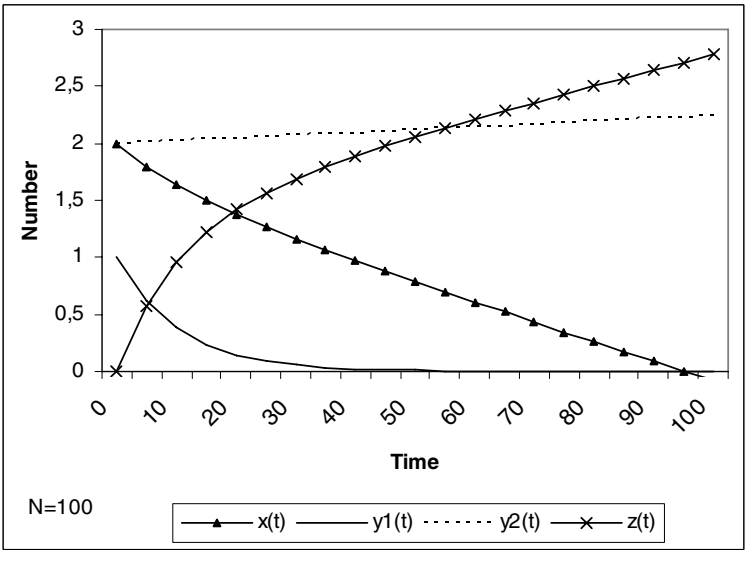
The Table 2 shows the second numerical simulation corresponding with the case when $R_{0,1} < 1$ and $R_{0,2} > 1$. We observe explicitly that $y_1(t)$ decays exponentially but $y_2(t)$ grows exponentially, which is a symptom of instability, and in this case there is partially developed outbreak. The Figure 1b) shows the corresponding epidemic curves and the typical behaviour of instability: the number of susceptible individuals is decreased to zero and the number of infected people grows exponentially, which means there is a partially developed outbreak.

Table 2. Simulations of epidemic thresholds and prevalence of clinical and subclinical cases according to the critical population size

Simulation	Critical population size N	$R_{0,1}$ (clinical cases)	$R_{0,2}$ (sub clinical cases)	$y_1(t)$ Prevalence of clinical cases	$y_2(t)$ Prevalence of sub clinical cases
1	10	0,019	0,112	$y_1(t)=e^{(-0,11*t)}$	$y_2(t)=2*e^{(-0,008*t)}$
2	100	0,195	1,123	$y_1(t)=e^{(-0,09*t)}$	$y_2(t)=2*e^{(-0,001*t)}$
3	1000	1,955	11,143	$y_1(t)=e^{(0,11*t)}$	$y_2(t)=2*e^{(0,10*t)}$

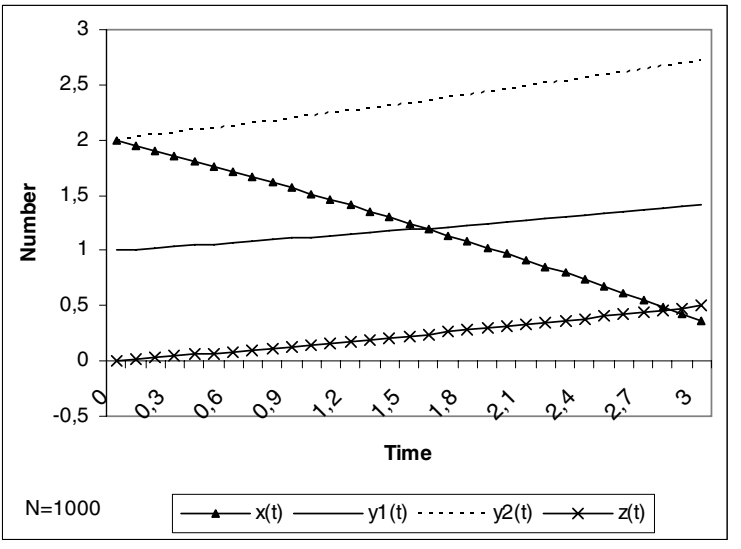


(a)



(b)

Fig. 1. Simulations of susceptible individulas *(x(t)),* clinical infected individuals *($y_1(t)$),* subclinical infected individuals *($y_2(t)$),* and removed individuals *($z(t)$)* by time, according to the critical population size: *a) N=10, b) N=100, c) N=1000.* Parameter values: Clinical infection rate (β_1)= 0,00025; subclinical infection rate (β_2)=0,00012; natality/mortality rate (μ)=0,00002; loss of immunity rate (q)=0,003; immunization rate (p)=0,0002; recovery rate of clinical cases (γ_1)=0,12; recovery rate of sub clinical cases (γ_2)=0,01.



(c)

Fig. 1. (*continued*)

Finally, the Table 2 shows the third numerical simulation corresponding with the case when $R_{0,1} > 1$ and $R_{0,2} > 1$.We observe explicitly that both $y_1(t)$ and $y_2(t)$ grow exponentially with the time, which is a sign of instability, and in this case there is a fully developed outbreak. The Figure 1b) shows the corresponding epidemic curves and the typical behaviour of instability: the number of susceptible individuals is decreased to zero and the number of infected people grows exponentially, which means there is a fully developed outbreak.

4 Discussion

The Table 1 shows that the simple SIR model only has one critical parameter, R_0. In contrast, according with Theorem 2, the SIIR model has two critical parameters, namely $R_{0,1}$ and $R_{0,2}$. It is a consequence of the introduction of two type of infected states: clinical and sub-clinical individuals. More over, the stability condition for the simple SIR model is merely $R_0 < 1$; but the stability condition for the SIIR model is more stringent because the Theorem 2 demands $R_{0,1} < 1$ and $R_{0,2} < 1$. The endemic states are more difficult to compute than the disease-free states. In general, computation of the endemic states demands the application of tools in computational commutative algebra and algebraic geometry. [10]

The epidemiology of sub clinical infections is largely unknown because there is not a reliable method to diagnose such infections, and follow-up studies about loss of immunity rate are scarce. However, from a theoretical point of view, studies about the effect of these sub clinical infections on the levels of infection, and the effect of waning and boosting of immunity on levels of infection in individuals with low (but

detectable) levels of immunity, who have experienced mild or sub clinical infections on contact with the virus, have been analyzed. [11], [12], [13].

The usefulness of this model is the theoretical illustration of two thresholds when considering clinical and sub clinical cases, although there are no real values of parameters for simulating the behavior of the disease with sub clinical infection. Simulation with rubella incidence in Latin America and the Caribbean in 1998 reflect the pattern of disease occurrence, although there are no data on infection rate for sub clinical infection over time. [9]

The algebraic expressions of the basic reproductive rate of the SIIR model give a synthesis of all epidemic parameters in the model and for this reason it is possible to appreciate the modifications of the basic reproductive rate when one or several epidemic parameters are altered, including cases when numerical values of such parameters are unknown and hard to obtain. It permits to derive control measures tending to reduce the basic reproductive rate, such as quarantine, surveillance, vaccination, education, sanitation, and so on.

This study describes the dynamics of the disease with two types of infected individuals but does not compare intervention strategies which could be useful especially when stochastic approaches of transmission in communities of households are considered. [6] However, it is observed that an epidemic with two type of infected people, according to a SIIR model, is more difficult to control than an epidemic ruled by the simple SIR model with only clinical infected individuals. Intensive contact tracing, syndromic surveillance and innovations in case detection could be required, when sub clinical and clinical infected individuals are considered. [4,6,13,14]

References

1. Brown, C., El Kahoui, M., Novotni, D., Weber, A.: Algorithmic methods for investigating equilibria in epidemic modelling. Journ. Symb. Comp. 41, 1157–1163 (2006)
2. Anderson, R.M., May, R.M.: Infectious diseases of humans: dynamics and control. Oxford University Press, New York (1992)
3. Diekmann, O., Heesterbeek, J.A.P.: Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation. John Wiley and Sons, New York (2000)
4. Chowell, G., Hengartner, N.W., Castillo-Chavez, C., Fenimore, P.W., Hyman, J.M.: The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. Journal of Theoretical Biology 229(1), 119–126 (2004)
5. Hsua, S.-B., Hsiehb, Y.-H.: On the Role of Asymptomatic Infection in Transmission Dynamics of Infectious Diseases. Bulletin of Mathematical Biology 70, 134–155 (2008)
6. Ball, F., Becker, N.: Control of transmission with two types of infection. Mathematical Biosciences 200, 170–187 (2006)
7. Águas, R., Gonçalves, G., Gabriela, M., Gomes, M.: Pertussis: increasing disease as a consequence of reducing Transmisión. Lancet Infect. Dis. 6, 112–117 (2006)
8. Bailey, N.T.J.: The mathematical theory of epidemics, p. 194. Charles Griffin and company limited, London (1957)
9. Panamerican Health Organization. Health conditions and trends. Health in the Americas, edn. Washington, DC, pp. 58–207 (2007)

10. Ospina, J., Hincapie, D.: Mackendrick: A Maple Package Oriented to Symbolic Computational Epidemiology. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3991, pp. 920–923. Springer, Heidelberg (2006)
11. Glass, K., Grenfell, B.T.: Waning immunity and subclinical measles infections in England. Vaccine 22, 4110–4116 (2004)
12. Glass, K., Grenfell, B.T.: Antibody Dynamics in Childhood Diseases: Waning and Boosting of Immunity and the Impact of Vaccination. J. theor. Biol. 221, 21–131 (2003)
13. Van Boven, M., Melker, H., Schellekens, J., Kretzschmar, M.: Waning immunity and sub-clinical infection in an epidemic model: implications for pertussis in The Netherlands. Mathematical Biosciences 164, 161–182 (2000)
14. Eames, K., Keeling, M.: Contact tracing and disease control. Proc. R. Soc. Lond. B 270, 2565–2571 (2003)

Test Power for Drug Abuse Surveillance

Jarad Niemi[1], Meredith Smith[2], and David Banks[1]

[1] Duke University, Durham, NC 27708 USA
[2] Purdue Pharma L.P., Stamford, CT 06901 USA
(now at Abbot Laboratories, Abbot Park, IL 60064 USA)

Abstract. Syndromic surveillance can be used to assess change in drug abuse rates and to find regions in which abuse is most common. This paper compares the power of three syndromic surveillance procedures (a paired-sample test, a process control chart, and a conditional autoregressive model) for detecting change in opioid drug abuse patterns, using data from two reporting systems (the OTP and PCC datasets). We find that the conditional autoregressive model provides good power and geographic information and that the OTP data carry the strongest signal.

1 Introduction

The substantial rise in nonmedical use and abuse of prescription opioid analgesics over the past decade offers an important opportunity for comparing syndromic surveillance methods [4]. Prescription drug abuse has similarities to infectious disease due to the inherent geographical effects [3], multiple reporting systems which vary in coverage and data quality, and it is a major concern in public health management. Prescription opioid analgesic abuse cost the U.S. an estimated $8.6 billion in 2001 due to increased health care, workplace, and criminal justice costs [1].

This paper contrasts three different strategies for syndromic surveillance:

- A paired-difference two-sample test, which looks for differences in abuse rates over time at each reporting site.
- A sequential process control procedure, using the CUSUM chart, similar to that used by the CDC [9].
- A conditional autoregressive (CAR) model which incorporates covariates as well as a model for geographic dependence.

These methods are compared with respect to their power in detecting simulated signal using historical abuse data and in their ability to detect hot spots of this abuse.

The data sets used in this study operate under the auspices of the Researched Abuse, Diversion, and Addiction-Related Surveillance (RADARS®) system:

- OTP. The Opioid Treatment Programs study collects quarterly questionnaires from abusers enrolled in Methadone Maintenance Treatment Programs (MMTPs) and thus captures a key population of sophisticated abusers.

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 131–142, 2008.

– PCC. The Poison Control Center network records information on help calls resulting from intentional drug exposures; not all poison control centers participate, but its coverage is about 70% of the U.S. by population.

Besides these databases we also considered: the National Survey on Drug Use and Health (NSDUH), an annual federal survey; Monitoring the Future (MTF), a nationally representative cohort study of self-reported drug use by 8th, 10th, and 12th grade students, who are then followed in biennial surveys until age 29; and the Drug Abuse Warning Network (DAWN), which provides an annual cross-sectional sample of Emergency Department visits related to nonmedical use of drugs. Although NSDUH and MTF survey 70,000 and 50,000 respondents respectively, the percentage of these individuals abusing opioids is small. Therefore their effective sample sizes for detecting change in a syndromic surveillance program would be inadequate. The effective sample size in DAWN is larger, but DAWN studies only a few major metropolitan areas in the country and therefore spatial information is limited.

We focused on one specific medication, the opioid analgesic OxyContin® (oxycodone HCI, controlled-release) Tablets, since it has been the target of abuse over several years [2]. We focus on change detection for a one-sided alternative which specifies that the drug abuse rate has decreased over time. This approach is simpler than two-sided alternatives, reflects federal interest in measuring the effectiveness of drug prevention programs, and our results extend directly to the symmetric hypothesis that drug abuse has increased.

All power studies were performed by simulation. For each combination of database and analysis, we examined power as a function of simulated levels of abuse reduction. The simulations were performed by bootstrapping [5] from the original data sets, after adjustment to achieve specified reduction levels.

Our goals are to determine the tradeoffs among the three analyses, in terms of power, geographic localization, and operational requirements (computing time, statistical complexity). We also want to determine the tradeoffs among the two databases used in this study. Section 2 describes the methodology; Section 3 presents the results; and Section 4 summarizes the comparisons.

2 Methodology

The methods developed in this paper follow this protocol: 1) a generative model is assumed for the data, 2) simulations with artificial signal are generated from this model, and 3) multiple surveillance techniques are applied to the resulting data. In the OTP data, the generative model is a CAR model including covariates. In the PCC data, the generative model is a log-linear model. In both data sets, the surveillance techniques used include a two-sample test and a process control chart. We also use the CAR and the log-linear model for surveillance in the OTP and PCC data sets, respectively, but the models include a term to detect the difference in abuse between time points.

2.1 Statistical Tests

Many statistical approaches to syndromic surveillance could be considered: repeated measures MANOVA, longitudinal analysis, time series analysis, various regression models, process control charts, two-sample tests, and so forth. In picking the three methods used in this paper, we sought transparency as well as adequate statistical power.

Two-Sample Tests. These tests look for a change between two time points. We use the classic one-sided test for a difference in binomial proportions. The test statistic is:

$$z = (\hat{p}_1 - \hat{p}_2)/\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $\hat{p}_1$ is the observed proportion of abusers in the previous quarter and $\hat{p}_2$ is the observed proportion in the current quarter. This test statistic is referred to a standard normal table.

The two-sample test can be improved when the same sites report each quarter. If there are k such sites, one can perform the two-sample test separately at each, and pool the resulting P-values according to Fisher's rule [6]. Let p_i, for $i = 1, \ldots, k$, be the P-value for site i; then

$$\chi^2_{2k} = -2\sum_{i=1}^{k} \ln p_i$$

which can be referred to a chi-squared table. Thus many slightly significant reductions can be pooled to give stronger evidence of a reduction. Repeated use requires adjustment for multiple testing. To give an indication of geographical variability in the reductions, one can map the P-values by region.

Process Control Charts. Control charts check whether a succession of observations has drifted away from a baseline value. A CUSUM chart plots the sum of the differences between the previous quarters' proportions and a baseline proportion. As described in [8], when this sum falls below a lower control line, the result is statistically significant. The CUSUM procedure is more complex than the two-sample test or the Shewhart control chart, but still fairly simple.

Control charts assume that the baseline is fixed and known. This is reasonable in manufacturing, but in syndromic surveillance, we do not know the baseline abuse rate; we can only estimate this, with uncertainty, from historical data. The assumption of no trend in this historical data is critical.

CAR Models. The previous tests make no use of covariate information. If covariates are influential, then a regression model should have more power and greater ability to notice specific regions with unusual behavior.

We use the Generalized Linear Model (GLM) in conjunction with the Conditional Autoregressive (CAR) model. Our GLM uses the logit function (log odds) to linearize the dependence of a proportion upon covariates [7].

CAR models include spatial or temporal dependence through a neighborhood structure, so that reporting units that are near each other have correlated data. This is reasonable in syndromic surveillance. In our application, there are known hot spots of opioid drug abuse in Appalachia and Maine. For this analysis, we aggregated the geographic information to the state level. The methodology could be extended to other spatial resolutions, but we found this aggregation to be effective for our purposes

Inference is done through Markov chain Monte Carlo [10], but in a problem of this scale there are computational challenges. Sometimes it took a full day of computing to provide a single point in the power curve.

2.2 Simulation Procedures

All of the power curve figures in Section 3 were produced by simulation. These simulations were produced using either the model in (1) or (2) where the mean was multiplied by the appropriate fraction to produce, on average, a linear decline in abuse over three years. The parameters used in these simulations were drawn from the posterior distributions for the parameters using only the pre-intervention data. We focus on power for two significance levels: $\alpha = .05$ and $\alpha = .05^2 = .0025$ that bracket loose and stringent levels for Type I error. Each plotted point is based upon 200 simulations with a specific, simulated decrease in abuse from the previous historical record in each of the databases.

The process control chart simulations assume that, after the last historical quarter, the abuse rate in subsequent quarters drops linearly over three years to a new level that was 5%, 10%, 15% or 20% lower. Extensive pre-simulation runs were made to estimate the lower control line values for these charts.

3 Results of the Power Analyses

The following subsections describe the power curves. There is a short review of each dataset and special analytic issues that they pose.

3.1 OTP

The OTP data derive from questionnaires administered to abusers enrolled in selected Methadone Maintenance Treatment Programs (MMTPs). It captures information on opioid abuse in the past 30 days, the primary drug, and geographic/demographic information. The data are quarterly in 2005.

OTP Two-Sample Test. Figure 1 shows the estimated power of the two-sample test using OTP data with Fisher's test for change at MMTP clinics that appear in both time periods. This "blocking" of an MMTP with itself automatically controls for many biases and reduces the variance in comparisons.

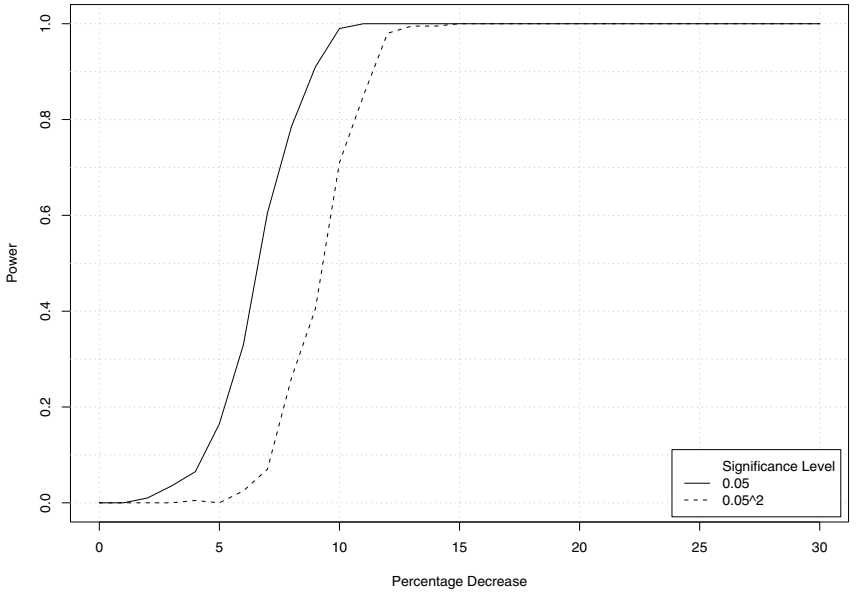


Fig. 1. Power curves for a two-sample OTP test. The most recent four quarters are combined to give the pre-sample abuse level.

OTP Control Chart. For control charts, one cannot plot power for all possible values of reduction (percentage decrease). So Figure 2 plots the probability of rejection for four different reduction levels: 20%, 15%, 10% and 5%, reading the curves from left to right.

Interpreting Figure 2 requires some care. Note that the two-sample tests use a year's worth of data, whereas the quarterly data has necessarily smaller sample size. Looking at the power in the fourth quarter gives a basis for comparison, but recall that control charts do not adjust for multiple testing.

OTP CAR Model. OTP data capture age, gender, race, and location. This enables use of a CAR model that incorporates spatial correlation structure.

The CAR model used in this power study is:

$$\begin{aligned} Y_{ik}(t) &\sim \text{Bernoulli}(p_{ik}(t)) \\ \text{logit}(p_{ik}(t)) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + b_k \end{aligned} \tag{1}$$

where $Y_{ik}(t)$ is the outcome for individual i living in state k (i.e., it is 1 if the individual has used an oxycodone product in the past 30 days and 0 otherwise). The covariates describe gender, race, and age, respectively, where race has been dichotomized to white or non-white and age has been broken down into 17 age categories. Flat priors were used for all coefficient parameters. The spatial random effect for U.S. state are assumed to have a CAR prior distribution

$$\pi(b_1, \ldots, b_K) \propto \exp\left(-\frac{\tau}{2}\sum_{i \neq j} w_{ij}(b_i - b_j)^2\right)$$

where τ is the precision parameter, given a Gamma(1,1) prior, and w_{ij} is obtained from the matrix of binary weights that indicate whether two states share a border. We examined the use of a quadratic term in age, but it was not significant and thus excluded from this model.

Table 1 shows the posterior credible regions (Bayesian confidence intervals) on the coefficients for gender, race, and age. The estimated value for the gender coefficient is .46 and its credible region excludes 0, so men are more likely to abuse opioid drugs. Similarly, the coefficient of 1.15 on race means that whites are more likely to abuse opioids. The age coefficient is negative, so the elderly are less likely to abuse.

The location terms b_k are random effects, and correlated. Since data are aggregated by state, the correlation structure is coarse: two states directly interact if they are contiguous; otherwise, states are conditionally independent given their neighbors. (Alaska is the only reporting state that had no neighbors.) The model for the correlation is multivariate Gaussian with unknown but common correlation for states that share boundaries.

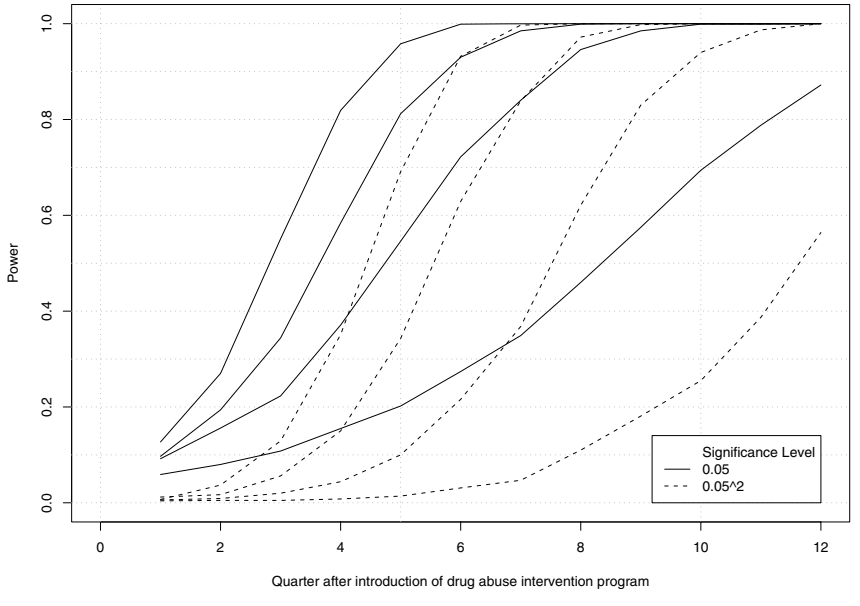


Fig. 2. Power curves for CUSUM testing with OTP data. The solid (broken) lines correspond to .05 (.0025) level tests. The lines, reading from the bottom up, correspond to 5%, 10%, 15% and 20% reductions in abuse rate, pro-rated over three years.

Table 1. Percentile points of the posterior distributions on the fixed-effect terms in the CAR model

	2.5%	50%	97.5%
Intercept	-2.2	-1.7	-1.2
Gender	0.31	0.46	0.58
Race	0.89	1.15	1.47
Age	-0.051	-0.043	-0.034
CAR precision	0.19	0.36	0.64

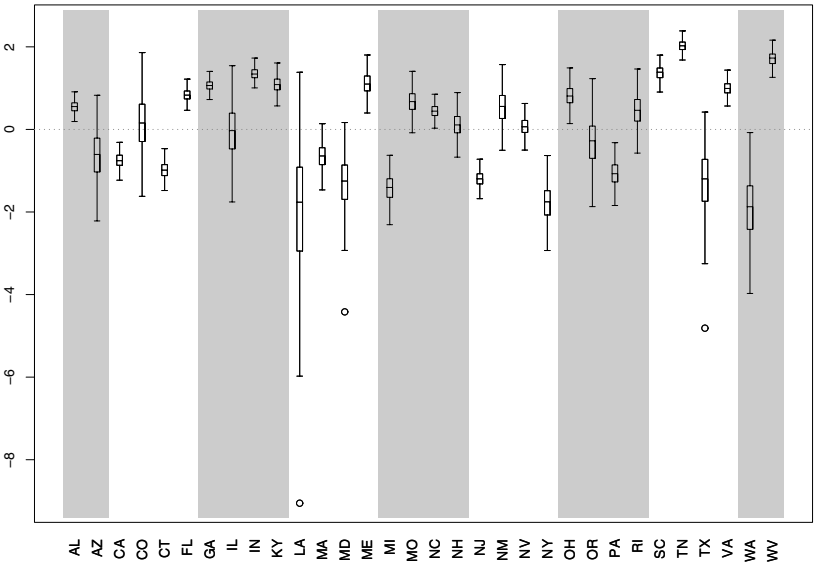


Fig. 3. A display, by state, of the 95% credible regions on the location (or state) effect. If the range of the line for a state straddles zero on the y-axis, then there is more than a 5% chance that that state has no effect on the abuse rate, after accounting for gender, race and age. The central boxes contain the middle 50% of the probability mass for the magnitude of the state effect, and the midline within the box is the point estimate of the magnitude of the state effect.

Figure 3 shows estimates of the state effects in the CAR model. Three states with large positive effects were Tennessee, West Virginia, and Virginia. This accords with previous reports of high opioid abuse rates in Appalachia. Some states, such as California and Connecticut, have lower than expected rates of opioid abuse. The wide interval for Louisiana surely reflects uncertainty in the data due to Hurricane Katrina.

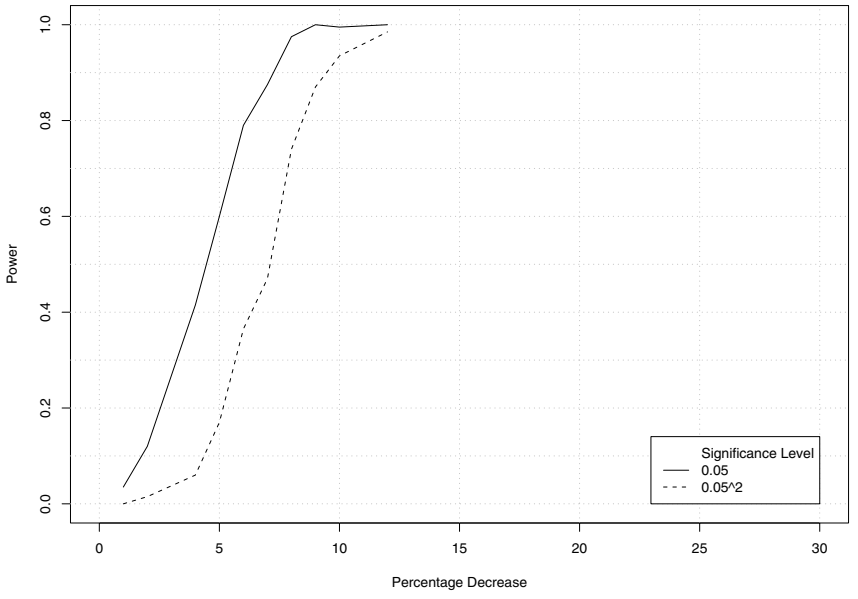


Fig. 4. Power curves for a CAR model test of abuse reduction using OTP data. The solid line is for an alpha level of .05; the dashed line is for an alpha level of .0025.

To test for a drop in abuse rate, the CAR model in equation (1) is modified to include a term for time. If the time coefficient is significantly less than zero, this indicates one-sided change (reduction). The magnitude of the effect can be estimated from the coefficient.

The CAR model was fit using Gibbs sampling run through WinBugs with an R interface. The OTP data takes a relatively long time to run (about 10 minutes per simulation run) so the power is only calculated for reductions of 1%, 2%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, and 12%, as shown in Figure 4.

3.2 PCC

The PCC network data consists of calls to regional poison control centers (usually funneled through 911 calls) regarding intentional exposure to drugs. The network covers approximately 70% of the U.S. population. The geographic counts are aggregated at the 3-digit zip code (3DZ) level. The data are highly reliable in terms of the identification of specific drugs, since PCC operators usually obtain the NDC code from the pharmacy label, but demographic data are not consistently captured.

The main explanatory variable in the PCC data is the 3DZ location. We use this to leverage information on the estimated population and the number of unique recipients of dispensed drugs (URDDs) in the region. The fitted model

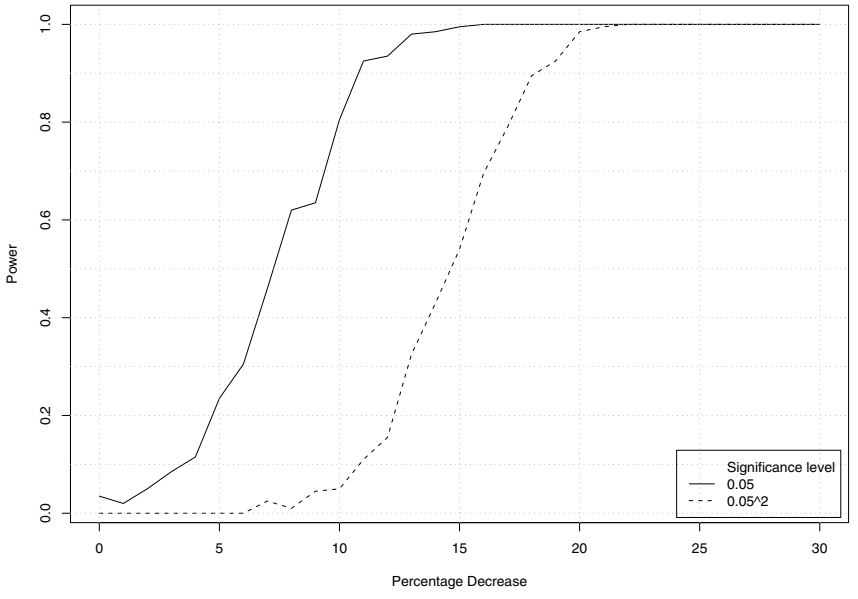


Fig. 5. The solid (broken) line is the probability of rejecting the null hypothesis of no reduction in opioid abuse at the .05 (.0025) level using PCC data with a two-sample test when, in fact, the reduction is as shown on the x-axis

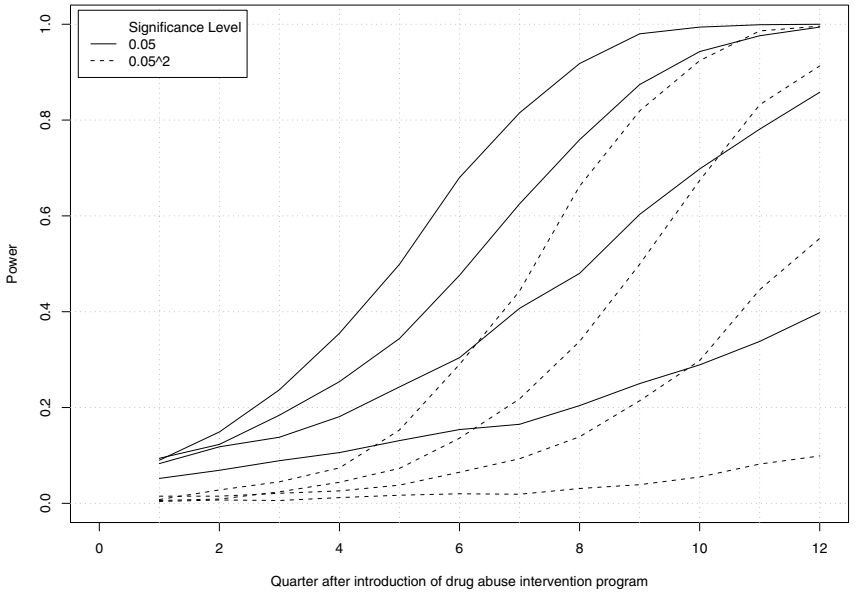


Fig. 6. The power curve for a CUSUM test with PCC data. The solid (broken) lines correspond to .05 (.0025) level tests. The lines, reading from the bottom up, correspond to 5%, 10%, 15% and 20% reductions in abuse rate, pro-rated over three years.

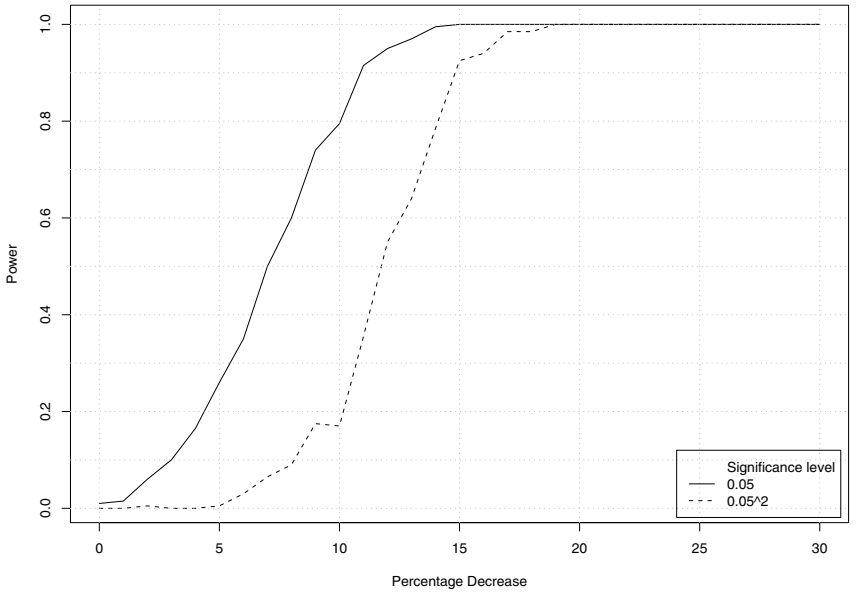


Fig. 7. The solid (broken) line is the probability of rejecting the null hypothesis of no reduction in opioid abuse at the .05 (.0025) level using PCC data and the GLM test when, in fact, the reduction is as shown on the x-axis

assumes the observed count Y_i of specific opioid PCC calls in the ith 3DZ has Poisson distribution with parameter λ_i given by:

$$\ln(\lambda_i) = \alpha \ln(\text{URDD}_i) + \beta \ln(\text{Pop}_i) \tag{2}$$

The population size and URDD provide rough "denominators" for the group at risk for opioid abuse.

PCC Two-Sample Test. After fitting, we used equation (2) to simulate observations for future quarters by drawing counts from Poisson distributions with decremented values for the estimated λ_i such that, on average, the number of simulated abuses decreased to a specified amount linearly over three years. The test statistic for deciding whether there has been a change is:

$$t_{n-1} = \bar{\Delta} / \sqrt{s_\Delta^2/(n-1)}$$

where n is the number of 3DZs and $\bar{\Delta}$ is the average regional difference in the count at the historical baseline of surveillance and the count afterwards. The s_Δ^2 is the sample variance of the observed differences.

In calculating power, we used only the 571 3DZs that reported in all quarters of 2005. Figure 5 shows the result.

PCC Control Chart. Using the simulation procedure described previously and thus taking advantage of information on URDD and population size, we generated data for the CUSUM test in Figure 6. The mean of the initial year was the baseline against which deviations should be discovered.

PCC Regression Model. We fit a GLM model using URDD and population size as covariates, with the addition of a term for time (as we did with the CAR model for OTP). The effort needed to map the 3DZs to a state-level adjacency matrix prevented a full CAR analysis, although this could be done. Figure 7 gives power curves for tests at the .05 and .0025 levels.

4 Conclusions

Disease surveillance seeks to identify a sudden increase in the prevalence of an illness against a background of relatively low rates. But drug abuse usually has a higher background prevalence which does not increase drastically. So drug abuse surveillance focuses on finding decreases in abuse that document successful intervention investments. This paper has compared methods and datasets that support that objective.

For OTP data, on a three-year horizon, the CUSUM is more powerful than the CAR test, which is more powerful than the two-sample test. But the CUSUM does not adjust for multiple testing, so its apparent power is misleading. Also, the three-year time frame gives it a larger effective sample size than the CAR or two-sample tests. For the PCC data, the same conclusions and caveats apply. The regression test is better than the two-sample test, but both lose to the CUSUM, which enjoys unfair advantages. (The CUSUM makes 12 tests, one for each quarter, all at a .05 level. So the overall probability of Type I error is actually $1 - (1 - .05)^{12} = .46$.)

In comparing the two data sets, OTP has a larger effective sample size than the PCC, so procedures that use OTP will generally be more powerful. Additionally, in this analysis, the regional information was more accessible and could be meaningfully interpreted for the OTP data.

People who contact a Poison Control Center are probably less sophisticated abusers than OTP clients. This may make them of greater (or less) public health interest. Data quality is also a issue. PCC data usually include the NDC code, but OTP data are self-reports from addicts based on recall. Other datasets (NSDUH, MTF, and DAWN) have similar data quality concerns.

References

1. Birnbaum, H.G., White, A.G., Reynolds, J.L., Greenberg, P., Mingliang, Z., Vallow, S., Schein, J., Katz, N.P.: Estimated costs of prescription opioid analgesic abuse in the United States in 2001: A societal perspective. Clinical Journal of Pain 22, 667–676 (2006)

2. Carise, D., Dugosh, K., McLellan, A.T., Camilleri, A., Woody, G., Lynch, K.G.: Prescription OxyContin abuse among patients entering addiction treatment. American Journal of Psychiatry 164, 1750–1756 (2007)
3. Cicero, T.J., Inciardi, J.A., Munoz, A.: Trends in the abuse of OxyContin and other opioid analgesics in the United States: 2002-2004. Journal of Pain 6, 662–672 (2005)
4. Compton, W.M., Volkow, N.D.: Major increases in opioid analgesic abuse in the United States: Concerns and strategies. Drug and Alcohol Dependency 81, 103–107 (2006)
5. Efron, B.: Bootstrap methods: Another look at the jackknife. Annals of Statistics 7, 1–26 (1979)
6. Fisher, R.: Combining independent tests of significance. The American Statistician 2, 30 (1948)
7. McCullagh, P., Nelder, J.: Generalized Linear Models. Chapman and Hall, London (1989)
8. Montgomery, D.: Introduction to Statistical Quality Control. John Wiley & Sons, Inc., New York (2001)
9. Rolka, H., Burkom, H., Cooper, G.F., Kulldorff, M., Madigan, D., Wong, W.-K.: Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: Research needs. Statistics in Medicine 26, 1834–1856 (2007)
10. Waller, L.A., Carlin, B.P., Xia, H., Gelfand, A.: Hierarchical spatio-temporal mapping of disease rates. Journal of the American Statistical Association 92, 607–617 (1997)

Assessing the Accuracy of Spatiotemporal Epidemiological Models

James H. Kaufman[1], Joanna L. Conant[2], Daniel A. Ford[1], Wakana Kirihata[3], Barbara Jones[1], and Judith V. Douglas[1]

[1] Healthcare Informatics Research, IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120, United States of America
[2] University of Vermont, College of Medicine, 89 Beaumont Avenue, Burlington, Vermont 05405, United States of America
[3] Wakana Kirihata, Columbia University, 3613 Lerner Hall, 2920 Broadway, New York, New York 10027, United States of America
kaufman@almaden.ibm.com, joanna.conant@uvm.edu, daford@almaden.ibm.com, wk2151@columbia.edu, bajones@almaden.ibm.com, judydouglas@comcast.net

Abstract. To demonstrate an approach that allows for the assessment of models and their accuracy, a numerical experiment was designed to generate a "control" data set and treated it as if it were "real" data. The open source spatiotemporal epidemiological modeler (STEM) was used to develop a control scenario depicting the spread of influenza in the state of Vermont; this scenario was then compared to three alternative models using such tools as root mean square differences and phase space analysis. This approach may prove helpful in responding to global pandemics and arriving at necessary policy decisions.

Keywords: Spatiotemporal data analysis, Infectious disease spread, Epidemiological models, Assessment, Validation.

1 Introduction

Any model of infectious disease is built on layers of assumptions and knowledge that define the underlying interactions between the infectious organism and the host species, and among members of the host community. At the highest level, the dynamics of disease spread in a human population depend on properties specific to the disease organism, such as its incubation, transmission, and mortality rate [1-3]. At the lowest level, independent of the disease organism, are the vectors that regulate the spread of the disease, such as the motion of people (traffic flow), the locations of waterways, even local rainfall (which in turn regulates mosquito populations) [1-3].

All of these interactions, disease specific and nonspecific, may be described as component models existing on a "graph" [4-6]. A graph is composed of nodes and edges connecting the nodes. Both the nodes and edges can have multiple labels

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 143–154, 2008.

defining properties. On a node, a label might represent rainfall at some location; on an edge, a label might represent the flow of people (or birds) between two locations. Labels and edges can also contain decorators that are like labels with built in algorithms to predict how properties might change with time. This is the abstraction used in the Spatiotemporal Epidemiological Modeler (STEM) [6], an open source framework for modeling infectious disease available through the Eclipse Foundation as part of its Open Health Framework [7].

1.1 Approaches to Assessment

Modeling such interactions as a graph allows models of infectious disease to be composed using layers of interchangeable and reusable parts [6]. The software architecture provided by STEM will allow scientists, across disciplines, to develop libraries of components to facilitate very rapid prototyping of new models in response to emerging infectious disease. These new capabilities bring with them the need for an approach to assess the relative accuracy of alternative models, and the loss in accuracy over time [8-10].

Infectious disease spread through a population is a stochastic dynamic process [9-12]. As in weather forecasting, any model of a "dynamical system" will lose accuracy at a rate dependant both on the uncertainty in the input data (knowledge of the disease state in the real population) and on the dynamics of the disease itself [9-12]. Some models, of course, do better than others. How does one assess accuracy of a model? Developing a "good model" depends not only on correct *tuning* of the model parameters (transmission rates, incubation rates, etc.), but also on choosing a model that correctly captures the *most essential vectors* of transmission. With modern computers it is tempting to build models that include as many conceivable interactions as possible. However, such models may be *over-determined* in that they introduce too many tuning parameters relative to the number of degrees of freedom in the input data set.

2 Developing a Control Scenario

The study of dynamical systems has yielded a variety of tools and formalisms that help scientists assess the accuracy of models. In this paper we present a control scenario built to test such an assessment. Using STEM, we generate a "control" data set for human influenza that we can treat as if it were "real" data. Our purpose is not to put forth the control data as a representation of any new disease, but rather to demonstrate an approach that allows for the assessment of models and their accuracy.

For the control scenario, we use the spread of human influenza within the state of Vermont. We chose influenza because it is a well studied infectious organism with well understood bounds on the disease parameters. For the base interaction layer, we chose Vermont because of its relatively small size, well defined geography, and characteristics of the state that restrict movement, as we explain below.

2.1 Methodology

In our approach, we generate the control data set and control scenario in four steps:

1) A "hidden" stochastic SEIR model is used to generate the control data set. This target data set logged as a comma separated variable (csv) file and treated as "real" disease data for the test.
2) The STEM Analysis Perspective, the component of the STEM disease modeling system that analyzes existing data sets, is then used to estimate the disease model parameters that most closely "fit" the "real data" generated in step 1. The unbiased estimation procedure in STEM is free of assumptions about the underlying spatial graph of interactions. It only averages over local epidemiological data. The following numerical experiment is also a test of the accuracy of this unbiased estimation procedure.
3) Using the parameter estimations in Step 2, several alternative scenarios for Vermont are composed using possible relationship graphs connecting the administrative level 3 (town) nodes.
4) Finally, the results of these alternative scenarios are compared to the control data by RMS difference and by a dynamical "Lyapunov" analysis in S-I space [13].

2.2 Basic Disease Model

The basic disease model identifies four states. A person can be susceptible (S), exposed (E), infectious (I), or recovered (R). Known as the SEIR model, it is defined by four equations [1-3]:

$$\frac{dS}{dt} = -\beta SI + \alpha R \tag{1a}$$

$$\frac{dE}{dt} = \beta SI - \varepsilon E \tag{1b}$$

$$\frac{dI}{dt} = \varepsilon E - \gamma R \tag{1c}$$

$$\frac{dR}{dt} = \gamma R - \alpha R \tag{1d}$$

Here β is the infection rate, α is the recovery rate, ε is the incubation rate, and γ is the immunity loss rate. Here we are simply using the most basic of compartment models for illustration purposes only. More sophisticated nonlinear models have been developed that capture richer dynamic behavior than the simple SEIR model [12].

A SEIR compartmental model assumes that everyone in the population is in one of the four states mentioned above. People in the susceptible state can contract the disease from other people who are infectious, while people in the exposed state have already contracted the disease but are not yet infectious. People in the recovered state are temporarily immune to the disease. For the purposes of this study both the birth rates and death rates are taken to be zero.

2.3 The Graph of Human Transportation in Vermont

Using publicly available geographical information system (GIS) data included in STEM [6], we compose our control scenario as a graph defining all administrative level 3 (town or zip code) regions of Vermont including human population and region areas. We use both the nearest-neighbor (adjacency) relationships for these regions and a graph of connections by roads. The graph defining the road network in Vermont is constrained by the Green Mountains (a sub-range of the Appalachian Mountains) that limit east-west road travel, increasing the likelihood that a disease will spread along major transportation corridors than to an adjacent town on the opposite side of the mountain range [14]. The edges in the road transportation graph that links towns together represent Interstate Highways, US Routes, and Vermont State Highways.

These edges must be weighted appropriately in order to capture the volume of traffic on the roads. This involves counting the number of roads crossing a particular border (connecting a pair of location nodes) and weighting each road connection by each class or capacity to carry traffic. Using annual average daily traffic data for a variety of locations, we calculate that, on average, Interstate Highways and US Routes are 3.2 times and 1.6 times more traveled than Vermont State Highways, respectively [15]. Thus, the weight of Interstate Highways, US Routes, and Vermont State Highways are 3.2, 1.6, and 1, respectively. The label defining total "relative" weight of an edge connecting nodes i and j is then given by:

$$W_{edge}(i,j) = 3.5N_{interstate}(i,j) + 1.7N_{USHwy}(i,j) + 1.0N_{VTRoute}(i,j) \quad (2)$$

If there are multiple connecting roads, the weights of the roads are summed to determine the total weight of that edge. For example, if two towns were connected by an Interstate Highway and a Vermont State Highway, the weight of this road would be 4.2. This un-normalized label defines the weight of the connection (in road edge units). It reflects the relative road capacity between sites i,j relative to other sites. The edge weight itself is then multiplied by an overall scaling factor that defined the net flux of people across that edge in a time period of one day as shown in Table 1.

Table 1. Target Model for Influenza in Vermont

Immunity Loss Rate (alpha)	0.003
Incubation Rate (epsilon)	0.2
Recovery Rate (gamma)	0.1
Transmission Coefficient (beta)	4.0
% Traveling per road edge unit	0.02
% Traveling per day by nearest neighbor edges (independent or roads)	0.001

Using the disease parameters in Table 1, a control data set of influenza data was generated for Vermont with a time interval of one day over a 405 day period. The simulation was "seeded" with one infectious individual located in Addison, Vermont, on day 1. The screenshot in Figure 1 shows the state of the resulting epidemic after 60

days. Since the control simulation (deliberately) used a spatial transmission dominated by roads, the state after the first 60 days clearly shows the expected anisotropic spread as the disease follows the most traveled paths for human transportation.

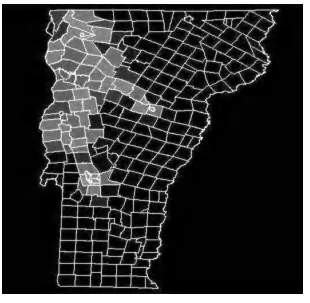


Fig. 1. Influenza in Vermont after 60 days with travel by both roads and nearest neighbors

3 Testing Accuracy and Value

If we then treat the control data set as "real" data of some epidemic, it is desirable to estimate the basic disease parameters or coefficients that would "best fit" the control data [16]. If accurate estimates can be obtained, one might expect to be able to model the time evolution of the epidemic. To accomplish this, we use the Parameter Estimation tools available in the STEM Analysis Perspective. Given any data set, this tool attempts to compute by a method of least squared fitting the basic parameter values for a standard SEIR model (STEM also provides SIR and or SI models; these are not tested here). The estimation enforces no underlying graph model for connections spatial regions in a data set; rather it assumes no connections. Estimations are made only for locations where an epidemic actually occurs. Given our foreknowledge of the correct answer (Table 1), the following numerical experiment is a first test for the accuracy and potential value of this approach to parameter estimation.

3.1 STEM Parameter Estimation

The four equations with four parameters that define the standard SEIR model are defined by equations 1a-d shown in Section 2.2. When rearranged by simple algebra, the four equations may be transformed as follows:

$$\frac{1}{I}\frac{d\ln S}{dt} = \alpha\frac{R}{SI} - \beta \tag{3a}$$

$$\frac{d\ln E}{dt} = \beta\frac{SI}{E} - \varepsilon \tag{3b}$$

$$\frac{d\ln I}{dt} = \varepsilon\frac{E}{I} - \gamma \tag{3c}$$

$$\frac{d\ln R}{dt} = \gamma\frac{I}{R} - \alpha \tag{3d}$$

All four of the above equations are in the form $y = mx + b$, where:

$$x = \frac{R}{SI}, y = \frac{1}{I}\frac{d\ln S}{dt} \tag{4a}$$

$$x = \frac{SI}{E}, y = \frac{d\ln E}{dt} \tag{4b}$$

$$x = \frac{E}{I}, y = \frac{d\ln I}{dt} \tag{4c}$$

$$x = \frac{I}{R}, y = \frac{d\ln R}{dt} \tag{4d}$$

Using the method of least squares, we can fit all four equations to a line and obtain both the slope (m) and intercept (b) from each equation. The slopes and intercepts in each equation correspond to disease parameters in the model equations.

The aim of this method is to estimate the parameters with the highest confidence level. That is, in addition to the standard deviations being the smallest, the equations are fit self-consistently and in the region for which they are most valid. To minimize the noise which arises from taking numeric derivatives of the discrete data, we form logarithmic derivatives for equations 4a-d.

Because of the divisors in the equations, the fits must be done in regions in which S, E, I, and R are all nonzero. Of course, there are variants of these equations which hold when one or more of these variables are equal to zero. An expanded version of the model, for future work, could utilize fitting in the regions of zero value as well. However, we find that these are relatively small regions of the data.

Moreover, we find that, although linear throughout the range of data, the equations gave regions of both positive and negative values for several of the parameters. We restricted our fit to the regions with positive values of the model parameters. This still allowed a significant range of times for the fitting.

The equations above give two independent estimated values for each parameter. To give equal weighting to all four equations, we averaged the two values for each parameter and calculated the standard deviations, as shown in Table 2.

Table 2. Parameter Estimation Results

Parameter	Estimated Value
Immunity Loss Rate (alpha)	0.001±0.005
Incubation Rate (epsilon)	0.219±0.005
Recovery Rate (gamma)	0.099±0.006
Transmission Coefficient (beta)	4.45±1.03

3.2 Candidate Models for Comparison

The values shown in Table 2 do not in themselves define an accurate model for the control scenario. The disease model itself must be constructed on top of a graph representing the spatial vectors that allow the disease to spread. Thus, we put forward three candidate models for the control scenario, using the estimated disease parameter values in each. In the first model we allow only disease transmission by roads with a weighting of 0.02 per road edge. In the second model we leave out the road edges

completely and only consider a graph with nearest neighbor connectivity and a flux of 5% of the population per nearest neighbor edges. In both cases the edges represent a symmetric bi-direction circulation between node edges per day (so there is no net population migration). Finally, in the third model we apply the same disease parameters using a "correct" model for transportation that includes both roads and nearest neighbors using the same weightings defined in the control scenario.

4 Results and Discussion

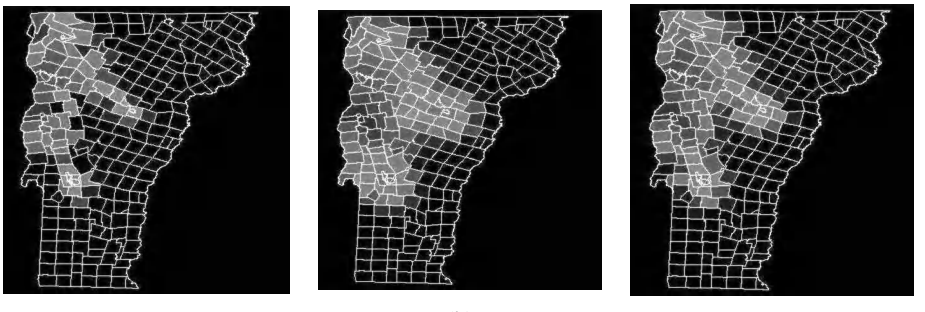
Figure 2 shows the state of the simulation at 60 days from the initial seed with each of the three transportation models defined in Table 3. These may be qualitatively compared with the reference scenario shown in Figure 1. To make a quantitative comparison we use two tools available in STEM's Analysis Perspective. The first tool measures the root mean square (RMS) difference $\delta_{a,b}(t)$ between two scenarios 'a' and 'b' at time t averaged over all locations (i). The algorithm used is:

$$\delta_{a,b}(t) = \frac{\sum_i P_i \sqrt{([I(a,t)_i - I(b,t)_i]^2 + [S(a,t)_i - S(b,t)_i]^2)}}{\sum_i P_i} \quad (5)$$

Note that each location '*i*' is weighted by the local population P_i so more populous locations contribute more to the measurement. The method assumes the scenarios a,b each cover the same set of location {i} which is true in this case. The RMS difference

Table 3. Three Models for Transportation

Candidate Scenarios	Fraction traveling per road edge unit per day	Fraction traveling per day by nearest neighbor edges per day
Roads Only	0.02	0.0
Nearest Neighbors Only	0.0	0.05
The "Correct" Graph including Both	0.02	0.01



(a) (b) (c)

Fig. 2. Screenshots of three models scenarios at 60 days with three models of transportation (a) Road Transport only, (b) Nearest Neighbor Edges only, and (c) both vectors included with the same weights as the control

is evaluated only at locations that have a non-zero rate of infection at some time t. If, for example, a comparison is made of two scenarios involving the entire planet but the epidemic in question is localized to a small region (e.g., the state of Vermont), one only wants to compare the model to the reference over those locations where the epidemic actually took place. Including locations with $I \equiv 0 \forall t$ would artificially reduce the estimated RMS error.

Finally we note that the comparison is made including only the S and I states. The algorithm compares these two disease state variables to allow for comparisons between different types of models. In a real epidemic scenario, one may wish to evaluate (for example) SI, SIR and SEIR models for the same infectious disease. Since the S and I states are common to all, measuring the RMS difference in this way allows for comparison across different types of models. In cases where only the infectious or perhaps change in infectious ΔI is known, RMS comparison may also be made based only on the I state.

Figure 3 shows the time dependent RMS error function for each of the three transportation models defined in Table 3 relative to the reference. The RMS difference as a function of time may begin near zero as a simulation is seeded or started at the same state as the reference data time series. Over time, as the spatiotemporal evolution of the model departs from the reference scenario, the RMS difference will increase. If the epidemic ends in both the model and reference scenarios, then the RMS difference will again approach zero. Averaging this error function in time provides a measure to the average difference or error of the model scenario relative to the reference. Figure 3 demonstrates that the integrated error is largest (7.5%) for the model scenario that assumed transportation of people based on a graph of nearest neighbor edges. A simple model using only road transportation had the lowest average error of 4.2%, lower even than the "correct" transportation model which had an average error of 4.3%. This is consistent with qualitative comparison of time evolution of the scenarios (comparing figures 2a-c with figure 1). Why should the simple road only transportation model outperform the "correct" transportation model when applied to the model scenario? As shown in Table 2, the estimated model parameters (compared with the reference in table 1) are inexact. Any disease parameter estimation procedures are likely to have sources of error. In this case, the estimation for the transmission coefficient (4.45) is high by slightly more than 10%. The average error measured by the RMS difference may be slightly lower in the road only transportation model; omitting the weak nearest neighbor connectivity used in the reference scenario somewhat compensates for the overestimate in the transmission coefficient.

The RMS error is a useful measure of the average difference between two scenarios. However, if a model and reference scenario each describe an epidemic that begins and end in the same state (zero infectious), the RMS error will eventually fall to zero even in case of a "bad" model. In addition to measuring the average error, it is useful to look for other measures that might provide a "fingerprint" for the spatiotemporal dynamics of an infectious disease. Like many dynamical systems, infectious disease is a process of many variables. However, it is often possible to capture the essential dynamics by looking at just a few system variables in an appropriate phase space. In its most general formalism, any dynamical system is defined by a fixed set of equations that govern the time dependence of the system's

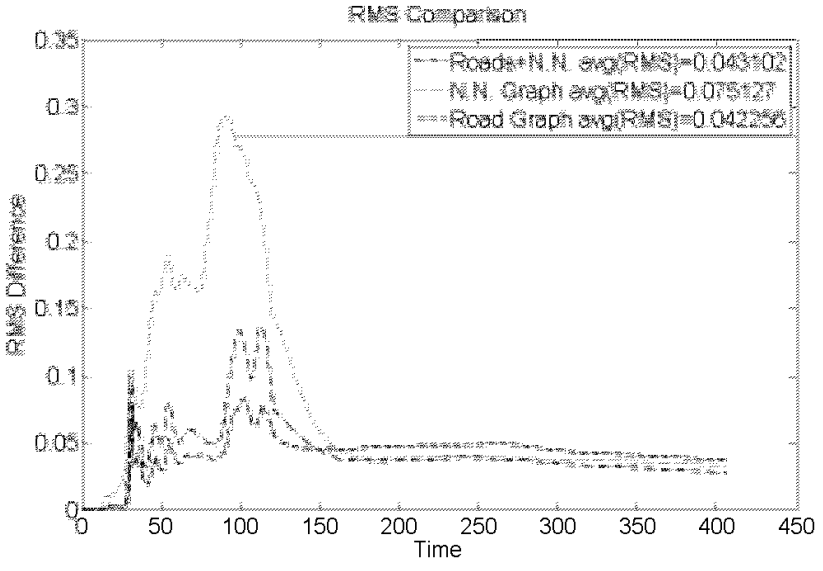


Fig. 3. RMS difference as a function of time between the three candidate scenarios and the reference Vermont data. Using the correct transportation model, the average RMS difference between the model and the reference is only 4.3% even with the errors in disease model parameters. A simple model using only roads has an average error of 4.2%. The nearest-neighbor only transportation model has the largest average error (7.5%).

state variables. The state variables at some instant in time define a point in phase space. A SEIR model defines a four dimensional phase space. An SI model defines a two dimensional space. Examining a reduced set of dimension may be thought of as taking slice through phase space (for example in the SI plane).

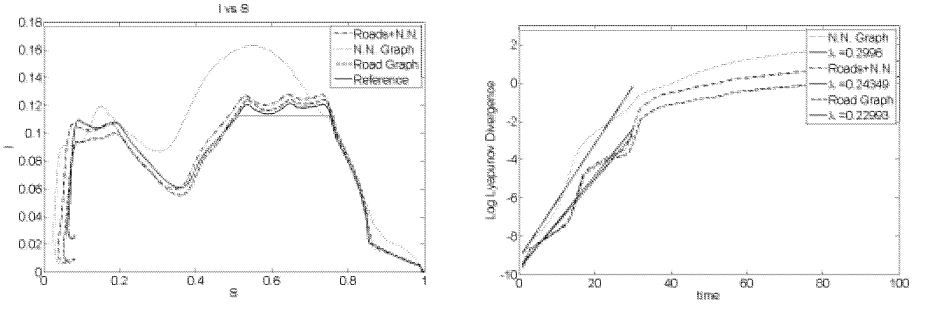
At the state of the system changes with time, the point (S(t), I(t)) in phase space defines a *trajectory* in the SI plane. Consider an epidemic that begins with one infectious person and virtually the entire population susceptible at t=0, S(0) ~ 1. The trajectory will begin at time zero along the S axis near 1. As the disease spreads, the susceptible population (S) will decrease and the infectious population (I) will increase. The detailed shape of the this trajectory will depend on the time it takes for the disease to spread to different population centers, as well as the (susceptible) population density function. The peaks and valleys along the trajectory in SI phase space proved a signature or fingerprint for an epidemic the shape of which depends on the disease, the disease vectors, the population distribution, etc. The mathematics of dynamical systems provide us with a formalism to compare trajectories in a phase space. Given a single set of rules (e.g., a disease model), two simulations that begin infinitesimally close together in phase space may evolve different in time and space. This separation in phase space can be measure quantitatively.

Vector $\vec{R}$ = (S(t),I(t)) defines a trajectory in SI space. The initial separation at time zero be defined as $|\delta\vec{R}_o|$. The rate of separation of two trajectories in phase space will often obey the equation

$$|\delta\vec{R}(t)| \approx e^{\lambda t} |\delta\vec{R}_o|$$

where λ is the Lyapunov Exponent. This exponent is a characteristic of the dynamical system that defines the rate of separation of infinitesimally close trajectories in phase space.

The STEM Analysis Perspective also provides tools to compare the trajectories describing the output of model scenarios with reference data sets (real data or other models). Figure 4a shows the trajectory in phase space for the reference scenario along with the three disease models based on the estimated parameters in Table 2 and the three models of transportation. The two models where transportation is dominated by road transport more accurately follow the reference trajectory (shown in black). The model based on nearest neighbor edges only has a clearly different trajectory fingerprint in phase space.



(a) (b)

Fig. 4. (a) The trajectories in Phase Space for the reference data (solid), and models with only road transport (broken), the correct transportation graph (uneven broken line), and a graph using nearest neighbor edges (gray). (b) The rate of separation of the three model scenarios from the reference scenario showing estimates for the Lyapunov exponent.

In Figure 4b, we plot the natural logarithm of the distance in phase space between each of the three model scenarios and the reference. The initial slope over the first 30 days on this semi-log provides a measure of the Lyapunov Exponent (λ). An accurate measure requires averaging over many such instances but the data in Figure 4 demonstrates the procedure and provides a qualitative comparison of the relative accuracy of each of the three models.

5 Conclusion

In response to any severe global pandemic, scientists will of necessity look toward modeling as a way to forecast the progress of the disease. Technology exists to rapidly build many new models of infectious disease. Clearly we need tools and procedures to compare the accuracy of competing models before we rely on any of them to guide policy [16,17]. The use of a control scenario described in this paper illustrates how tools available today may be tested in a pandemic exercise. Only by testing the accuracy of our models and our tools can we develop a useful framework

for the future. Because infectious disease evolves in time and space as a stochastic dynamical system, no model will ever be completely accurate.

5.1 Using Modeling to Test Policies

If we can develop "good approximations" and understand the rate at which models lose accuracy, then, much like the 5 day forecast in weather prediction, we can use modeling to test policies [17,18]. A "good approximation" requires not only knowledge of the disease parameters themselves, but also an understanding of the most important disease vectors and the "denominator data" that captures the transmission dynamics of the disease. In this exercise we estimate the disease coefficients locally by assuming a particular underlying graph of spatial connections. We then applied these estimated coefficients to models built on different spatial connection graphs. In the future it may be possible, with sufficient experimental data, to create algorithms that not only provide estimates for disease coefficients but also for the dominant transportation graph.

5.2 Directions for Future Research

A natural extension of the current estimation algorithms would allow communication between neighboring geographies, using roads, nearest neighbors, or both. The equations then become dependent on a large number of variables, and the analysis correspondingly becomes more complex. They can still be transformed into linear equations, however, and analyzed using the approach described in this paper.

A validated model such as the one described here could be used to confidently investigate and simulate a variety of infectious disease outbreaks. The results of these simulations could be used by healthcare organizations, public health officials, and governmental policy makers to understand how a disease may spread, and to determine appropriate ways to deal with such an outbreak.

This validation will require a variety of tools comparing model scenarios. Root mean square (RMS) differences and Phase Space Analysis can be applied to compare the relative accuracy of competing models and to estimate the rate at which any of these models lose accuracy. The same tools, applied to individual models, can also be used to measure the sensitivity of those models to initial conditions and, therefore, sensitivity to uncertainty in initial experimental data.

Acknowledgements

This project is being developed under Contract Number FA7014-07-C-0004, with the U.S. Air Force Surgeon General's Office (AF/SG) and administered by the Air Force District of Washington (FDW). The Air Force has not accepted the products depicted and issuance of a contract does not constitute Federal endorsement of the IBM Almaden Research Center. The authors thank Justin Lessler, at Johns Hopkins Bloomberg School of Public Health, for his input to the project, and acknowledge Eclipse for its support of the Open Health Framework (OHF) upon which STEM runs.

References

1. Hethcote, H.W.: The Mathematics of Infectious Diseases. SIAM Rev. 42, 599–653 (2000)
2. Bailey, N.T.: The Mathematical Theory of Infectious Diseases, 2nd edn. Charles Griffin and Co. Ltd., London (1975)
3. Anderson, R.M., May, R.M.: Infectious Diseases of Humans. Oxford University Press, Oxford (1992)
4. Widgren, S.: Graph Theory in Veterinary Epidemiology – Modelling an Outbreak of Classical Swine Fever. Thesis, Institution for Ruminant Medicine and Veterinary Epidemiology, Swedish University of Agricultural Sciences, Uppsala, Sweden (2004)
5. Myers, L.A., Newman, M.E.J., Martin, M., Schrag, S.: Applying Network Theory to Epidemics: Control Measures for Mycoplasma Pneumonia Outbreaks. Emerg. Infect. Dis. (2003), http://www.cdc.gov/ncidod/EID/vol9no2/02-0188.htm
6. Ford, D.A., Kaufman, J.H., Eiron, I.: An Extensible Spatial and Temporal Epidemiological Modeling System. Int. J. Health Geogr. 5(4) (2006), http://www.ij-healthgeographics.com/content/5/1/4, http://www.eclipse.org/ohf/components/stem/
7. The Eclipse Foundation, http://www.eclipse.org/org/
8. Schaffer, W.M., Bronnikova, T.V.: Parametric Dependence in Model Epidemics. J. Biol. Dynam. 1, 183–195 (2007)
9. Olsen, L.F., Schaffer, W.M.: Chaos vs. Noisy Periodicity: Alternative Hypotheses for Childhood Epidemics. Science 249, 499–504 (1990)
10. Ohkusa, Y., Sugawara, T.: Application of an Individual-Based Model with Real Data for Transportation Mode and Location to Pandemic Influenza. J. Infect. Chemother. 13(6), 280–389 (2007)
11. Kuznetsov, Y., Piccardi, C.: Bifurcation Analysis of Periodic SEIR and SIR Epidemic Models. J. Math. Biol. 32, 109–121 (1994)
12. Li, M.Y., Muldowney, J.: Global Stability for the SEIR Model in Epidemiology. Math. Biosci. 125, 155–164 (1995)
13. Zhang, G., Liu, Z., Maa, Z.: Generalized Synchronization of Different Dimensional Chaotic Dynamical Systems. Chaos, Solitons & Fractals 32(2), 773–779 (2007)
14. Wilbur Smith Associates: Northeast CanAm Connections – Integrating the Economy and Transportation. Task force report (2007)
15. Vermont Agency of Transportation, Policy and Planning Division, Traffic Research Unit: Continuous Traffic Counter Grouping Study and Regression Analysis Based on 2007 Traffic Data (2008)
16. Wearing, H.J., Rohani, P., Keeling, M.J.: Appropriate Models for the Management of Infectious Diseases. PLoS Med. 2,7, 621 (2005)
17. Bootsma, M.C.J., Ferguson, N.M.: The Effect of Public Health Measures on the 1918 Influenza Pandemic in U.S. Cities. PNAS 104, 18 (2007)
18. Arino, J., Frauer, F., van den Driessche, P., Watmough, J., Wu, J.: A Model for Influenza with Vaccination and Antiviral Treatment. J. Theor. Bio. 253(1), 118–130 (2008)

Simulation of Multivariate Spatial-Temporal Outbreak Data for Detection Algorithm Evaluation

Min Zhang, Xiaohui Kong, and Garrick L. Wallstrom

Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, 15260
{miz14,xik2,glw6}@pitt.edu

Abstract. We developed a template-driven spatial-temporal multivariate outbreak simulator that can generate multiple data streams of outbreak data for evaluating detection algorithms used in disease surveillance systems. The simulator is controlled via intuitive parameters that describe features of the outbreak and surveillance system such as the elevated risk of disease, surveillance data coverage, case behavior probabilities, and the distribution of behavior times. We provide examples of temporal and spatial-temporal outbreak simulations. Our simulator is a useful tool for evaluating of outbreak detection algorithms.

Keywords: Outbreak simulation, multivariate biosurveillance data.

1 Introduction

Biosurveillance systems collect and automatically analyze various types of data in search of signs of possible disease outbreaks. For example, a system may obtain, from a subset of emergency departments (EDs) in a region, the daily count of the number of ED visits for each of several syndrome categories. The system may also obtain data from laboratories, pharmacies, and other data providers. The automatic analysis of biosurveillance data is conducted by outbreak detection algorithms. Researchers have developed numerous algorithms for detecting outbreaks, ranging from temporal algorithms that detect changes in single time series for a single geographic region [1], spatial-temporal algorithms that utilize both spatial and temporal data [2], to algorithms that work on multivariate data streams [3].

Evaluation of detection algorithms requires surveillance data from outbreak and non-outbreak periods. Data are often available for non-outbreak periods. However, due to the rarity of real outbreaks, many evaluations cannot use data from real outbreaks. In such situations, researchers construct semi-synthetic data by simulating captured outbreak cases (those captured by the surveillance system) and adding those cases to real non-outbreak data.

One type of outbreak simulator consists of disease-specific simulators. For example, Hogan et al. [4] used a model of an aerosol anthrax release to evaluate detectability of anthrax outbreaks. Watkins et al. [5] have developed software for a geographic information system (GIS) environment for simulating spatial-temporal disease outbreaks. Such simulators can offer good face validity and permit investigation in to the role of meaningful outbreak parameters.

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 155–163, 2008.

Another type consists of simulators that are not disease-specific. These simulators are typically defined by a template function that describes the temporal shape of the outbreak in surveillance data. For example, Reis et al. [6] created outbreaks in temporal data consisting of 20 cases per day for 7 days. Researchers can tailor the shape, magnitude and duration to match those of a hypothetical outbreak of interest, and add noise to improve realism. Researchers have also used simple extensions of this approach to create outbreaks in spatial-temporal data [7]. Cassa et al. [8] developed the open-source AEGIS Cluster Creation Tool for simulating spatial-temporal outbreaks that are not disease-specific.

The above simulators all create temporally-aggregated counts, for example, daily counts of ED visits. The simulation of only aggregated data can create difficulties when comparing algorithms that run on data that are aggregated differently, or algorithms that run at different frequencies. While these difficulties are surmountable by using the least common denominator for temporal aggregation, or by simulating visit times in a second simulation step, an alternative solution is to simply simulate visit times directly instead of simulating aggregated counts. Zhang and Wallstrom adopted this approach and developed a template-driven outbreak simulator that generates event time data for a single data source [9]. This approach has the added advantage of forcing the logical separation between the simulation method and the aggregation routines used by surveillance systems.

One potential limitation is shared by those existing non-disease specific outbreak simulators mentioned above: they only simulate outbreak data for a single data stream (i.e. ED respiratory visits). Many biosurveillance systems monitor multiple data streams, most notably ED visits for multiple syndromes [10], over-the-counter (OTC) sales data for multiple product categories [11], and laboratory data [12, 13]. Semi-synthetic evaluation of algorithms that utilize multiple data streams requires outbreak simulations that simultaneously affect the multiple data streams.

In this paper, we introduce a non-disease specific outbreak simulator that creates multivariate outbreak data. While the simulator is largely based upon the univariate simulator developed in Zhang and Wallstrom, it is not a strict extension. It does, however, inherit many of its features such as the ability to generate purely temporal and spatial-temporal event time data in accordance with user-defined template functions. Our objective is to create a simulator of multiple series of event times that can be controlled intuitively through the use of template functions. We describe our multivariate outbreak data simulator in the context of OTC sales data for a particular category of products and ED visit data for a particular syndrome to illustrate the ease and flexibility of our simulator by generating temporal and spatial-temporal outbreaks datasets for both OTC sales and ED visit during the same outbreak period.

2 Methods

To simulate multivariate outbreak data, we develop a model that describes the effect of an outbreak on multiple streams of data. We begin by associating each data stream with a behavior. For example, we associate a data stream of ED respiratory visits with the behavior of visiting an ED with a respiratory chief complaint. Similarly, we

associate a data stream of OTC purchases of cough and cold products with the behavior of purchasing a cough and cold product from a pharmacy.

A rough outline of our simulation process follows:

Step 1: Determine the total number of outbreak cases. This number includes those cases not captured by the biosurveillance system.

Step 2: Distribute the cases geographically into regions. Consistent with common terminology, we call the geographic regions *tracts*.

Step 3: Determine the data streams that are affected by each case.

Step 4: Determine the event time for each case and affected data stream.

We first consider the purely temporal simulation of multivariate outbreak data. We then proceed to discuss spatial-temporal simulation of multivariate outbreak data.

2.1 Multivariate Purely Temporal Simulation

We discuss here the purely temporal simulation of M data streams. We first describe the inputs necessary for the simulation, and then describe how these inputs are used to carry-out the above simulation steps.

Outbreak Magnitude. This parameter, denoted by C, is the total number of outbreak cases, including those not captured by the biosurveillance system.

Behavior Probability Vector. The behavior probability vector is a vector of length 2^M consisting of the joint probabilities for the M behaviors for each outbreak case. For example, consider the simulation of a gastrointestinal outbreak with two data streams: ED gastrointestinal (GI) visit data and OTC antidiarrheal sales. A joint probability vector (0.1, 0.25, 0.35, 0.3) means that for any outbreak case, with probability 0.1 the case both went to an ED with a GI chief complaint and bought an antidiarrheal OTC product, with probability 0.25 the case went to the emergency room but didn't buy an OTC product, with probability 0.35 the case didn't go to an ED but bought an OTC product and with probability 0.3 the case neither went to an ED nor bought an OTC product.

Coverage Vector. The coverage vector consists of M coverage probabilities, which indicate the probability that a behavior is captured by the biosurveillance system in the outbreak region.

Temporal Template. The temporal template is a function f that describes how the rate of outbreak-related events changes across time. Specifically, we define f to be a joint density function for a vector $t = (t_1, t_2, \ldots, t_M)' \in [0, T_1] \times [0, T_2] \times \cdots \times [0, T_M]$ of event times, with one event time for each data stream. We decompose f into a product of conditional density functions:

$$f(t) = f_1(t_1) * f_2(t_2 \mid t_1) \ldots * f_M(t_M \mid t_1, t_2, \ldots t_{M-1}). \quad (1)$$

For convenience, we restrict f_i $(1 \le i \le M)$ to be a bounded (conditional) density function on $[0, T_i]$ $(1 \le i \le M)$.

With the above inputs, the simulation is carried out as follows:

Step 1: The total number of outbreak cases is given by the outbreak magnitude *C*. We note here that *C* may be either a fixed number, or could itself be randomly drawn from, say, a Poisson distribution with a fixed mean.

Step 2: This step is not necessary for purely temporal simulation.

Step 3: For each case we use the behavior probability vector to determine the behaviors that the case engages in. For each such behavior, we then use the associated coverage probability to determine whether that behavior is captured by the biosurveillance system.

Step 4: For each case, Step 3 determines the collection of captured behaviors for the case. We then simulate the vector of behavior event times by drawing an observation from the joint marginal distribution of event times for those behaviors.

Several assumptions are critical in this simulation process. First, we assume that each case is represented at most once in each data stream. Second, we assume that each event time is independent of *whether* other behaviors are engaged in and captures, but not necessarily independent of the times of such behaviors. For example, the marginal distribution of the ED visit time does not depend on whether the case purchases an OTC product or whether that purchase is captured by the system, but the ED visit time and OTC purchase time may be correlated. Finally, observe that due to this assumption, we can greatly simplify the simulation process by simulating the complete vector of event times for each case, and simply censoring those event times for behaviors not engaged in or not captured by the biosurveillance system.

Example. We illustrate temporal multivariate outbreak simulation by simulating bivariate data consisting of one OTC data stream and one ED data stream. We specify a template function that is linearly increasing for the OTC event time t_1, and increasing non-linearly for the ED event time t_2 conditional on t_1. Specifically,

$$f_1(t) = \begin{cases} 2t_1 / T^2 & 0 \le t_1 < T \\ 0 & otherwise \end{cases} \qquad (T = 5) \tag{2}$$

and

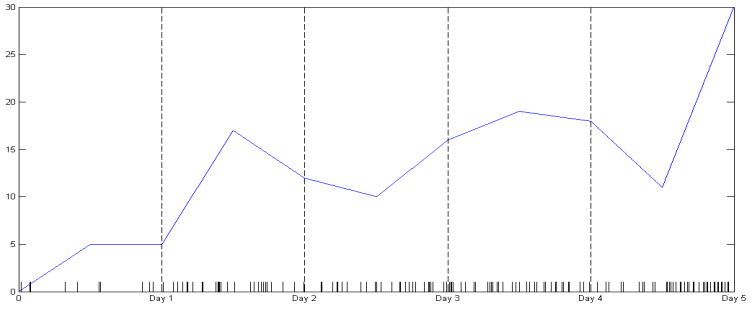
$$f_2(t_2 \mid t_1) = 0.25 \times (2t_2 / T^2) + 0.75 \times (1/(T - t_1)) \times I[t_1 < t_2 < T], \text{ where}$$
$$I[t_1 < t_2 < T] = 1 \text{ only if } t_2 \text{ is greater than } t_1 \text{ and 0 otherwise} \tag{3}$$

The distribution of $t_2 \mid t_1$ is a mixture: with probability 0.25, t_2 is distributed identically to t_1 and with probability 0.75, $t_2 \mid t_1$ is uniform on (t_1, T). The mean of the OTC event time is 3.33 days and the mean of the ED event time is 3.96 days. Our intuition to use this distribution in this example is to simulate the effect that ED event times are often later than OTC event times, which would occur if cases tend to

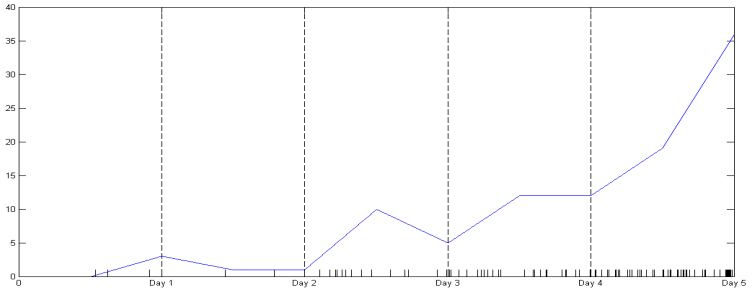
self-treat with OTC products earlier in the course of an illness when symptoms are less severe, and visit an ED later in the course of an illness when symptoms are more severe.

We simulate the multivariate outbreak that consists of OTC event times and ED event times. For each series, we set $C = 300$ cases, the behavior vector is (0.1, 0.1, 0.2, 0.6), and the coverage vector is set to (.70, .40) for OTC and ED respectively.

The simulated visit times are aggregated into 12-hour-aggregated counts and graphed in Figure 1 (a - b).



(a) OTC data stream



(b) ED data stream

Fig. 1. Simulated OTC (a) and ED (b) visit times using a semi-linear template function. 12-Hour-aggregated visit times are created for each time series.

2.2 Spatial-Temporal Multivariate Outbreak Simulation

We now turn to spatial-temporal simulation. The inputs for the simulation are similar to those for the purely temporal simulation. The main difference is that we replace the temporal template with a joint spatial-temporal template. We also extend the definitions of the other inputs.

Outbreak Magnitude. The outbreak magnitude C is the total number of outbreak cases across all tracts in the set of tracts S over the duration of the outbreak.

Behavior Probability Vector. The behavior vector here has the same definition as in our purely temporal version. The behavior probabilities could vary across tracts, or one could assume that behavior rates are the same across tracts (e.g. the likelihood of going to an ED is the same for every zip code in the region).

Coverage Vector. The coverage vector here has the same definition as in the temporal version. Like the behavior probability vector, the coverage can be a function of space or could be assumed to be uniform across the region.

Spatial-Temporal Template. The spatial-temporal template is a function f of time and space that describes how the rate of new cases changes across time and space. Specifically, we define $f(s,t) = f_S(s) f_T(t \mid s)$ to be a bounded joint probability mass function and probability density function over the spatial location and event times for each case. We interpret $f_S(s)$ as the probability that a case is assigned to tract s. This probability is a function of the elevated risk in tract s. Specifically, for tract s, let r_s denote the elevated disease risk , and n_s denote the population. Then

$$f_S(s) = \frac{n_s r_s}{\sum_{s' \in S} n_{s'} r_{s'}} \tag{4}$$

With the above inputs, spatial-temporal simulation is carried out as follows:

Step 1: The total number of outbreak cases is given by the outbreak magnitude *C*.

Step 2: Each case is assigned to a tract randomly according to the spatial template f_S.

Step 3: For each case we use the behavior probability vector, which may depend on the case's tract, to determine the behaviors that the case engages in. For each such behavior, we then use the associated coverage probability, which also may depend on the case's tract, to determine whether that behavior is captured by the biosurveillance system.

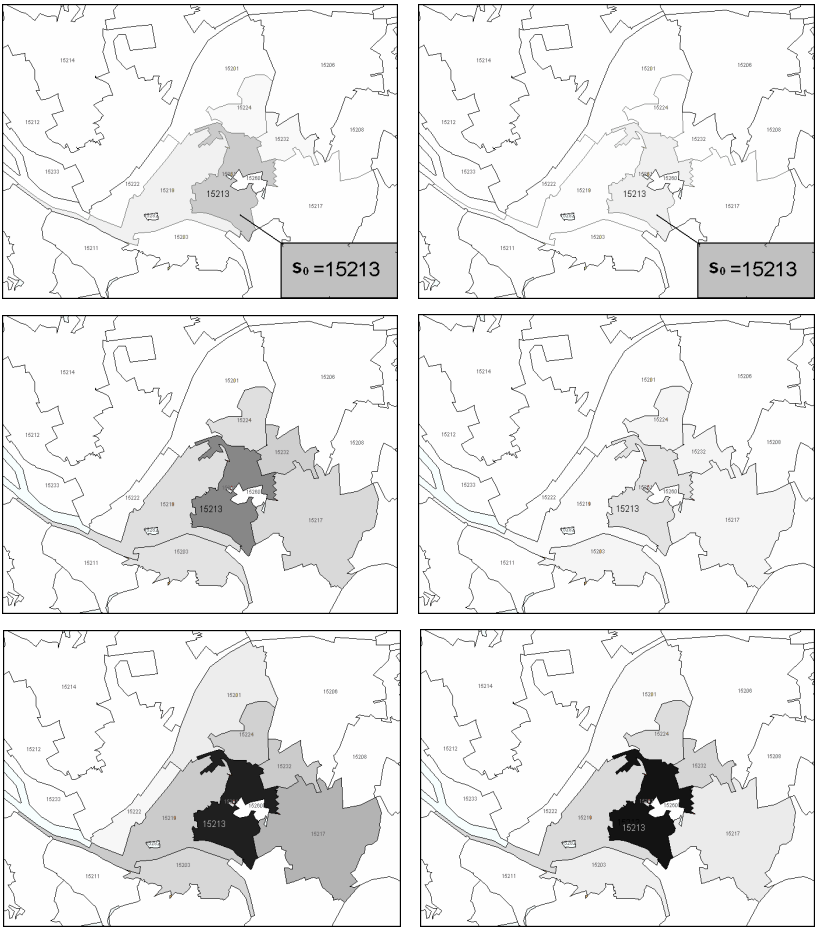
Step 4: For each case, Step 3 determines the collection of captured behaviors for the case. We then simulate the vector of behavior event times by drawing an observation from the joint marginal distribution of event times for those behaviors in that case's tract.

The assumptions listed for the purely temporal simulation remain in effect. In particular, one can simplify the simulation process by sampling the full vector of event times in Step 2 and censoring those for behaviors not engaged in or not captured by the biosurveillance system.

Example. In this example, we generate outbreak event times with a tract-dependent lag. Specifically, we assume that $f_T(t \mid s) = f^*(t - l_s)$ where l_s denotes the lag in the data stream in tract s. We define f_S to be a decreasing function of distance from the zip code s_0 = 15213 in the Pittsburgh area, with $r_s = 0$ for zip codes at least 7.4 kilometers from s_0. We set $C = 1200$ cases, $T = 3$ days, the behavior vector is (0.1,

0.1, 0.2, 0.6), and the coverage vector is set to (.70, .40) for OTC and ED respectively.

Figure 2 (a - b) shows the aggregated number of cases in the simulated outbreak in each affected zip code that were captured in the OTC and ED data streams respectively during each day:



(a) OTC visit times (b) ED visit times

Fig. 2. (a)~ (b) show the spatial-temporal OTC and ED data streams. Each sub-figure plot the daily aggregated counts of events in that area for each day (Top, day 1; Middle, day 2; Bottom, day 3). A lighter color indicates a smaller number of cases, while a darker color indicates a larger number of cases.

3 Discussion

We presented a simulator that researchers can use to generate visit times for multiple data streams across spatial tracts. These visit times can be injected into baseline data

to create semi-synthetic outbreaks that can be used to assess the sensitivity and timeliness of outbreak detection algorithms that utilize multivariate data.

To simulate multivariate data, we represent the dependency between different data streams with a behavior vector and a joint distribution over the event times. We also utilize a coverage vector to represent a critical property of a biosurveillance system. In particular, this enables evaluating a system with coverage probabilities that depend on both the data stream and the tract.

There is another sense in which data could be considered to be multivariate. Suppose for each case, demographic or other patient characteristics are also observed. It is straightforward to extend the present method to generate this kind of data as well. Specifically, one could specify a distribution over a set of demographic groups. For each case, the distribution could be used to determine the demographic group that the case belongs to. The spatial-temporal template, behavior vector, and coverage vector, which could depend on the demographic group, would be used to simulate the spatial location of the case and the event times for the captured behaviors that the case engages in.

Acknowledgements. This research was supported by a grant from the Centers for Disease Control and Prevention (R01PH000025). This work is solely the responsibility of its authors and do not necessarily represent the views of the CDC.

References

1. Wong, W.-K., Moore, A.W.: Classical time-series methods for biosurveillance. In: Wagner, M.M., Moore, A.W., Aryel, R.A. (eds.) Handbook of biosurveillance, pp. 217–234. Academic Press, New York (2006)
2. Lawson, A.B., Kleinman, K.: Spatial & syndromic surveillance for public health. John Wiley & Sons, West Sussex (2005)
3. Kulldorff, M., et al.: Multivariate scan statistics for disease surveillance. Statistics in Medicine 26(8), 1824–1833 (2007)
4. Hogan, W.R., et al.: The Bayesian aerosol release detector: an algorithm for detecting and characterizing outbreaks caused by an atmospheric release of Bacillus anthracis. Statistics in Medicine 26(29), 5225–5252 (2007)
5. Watkins, R.E., et al.: Using GIS to create synthetic disease outbreaks. BMC Medical Informatics & Decision Making 7, 4 (2007)
6. Reis, B.Y., Pagano, M., Mandl, K.D.: Using temporal context to improve biosurveillance. Proceedings of the National Academy of Sciences of the United States of America 100(4), 1961–1965 (2003)
7. Daniel, B.N., et al.: Detection of emerging space-time clusters. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, Chicago (2005)
8. Cassa, C.A., et al.: A software tool for creating simulated outbreaks to benchmark surveillance systems. BMC Medical Informatics & Decision Making 5, 22 (2005)
9. Zhang, M., Wallstrom, G.L.: Template Driven Spatial Temporal Outbreak Simulation for Detection Algorithm Evaluation. In: AMIA Annual Symposium Proceedings/AMIA Symposium 2008 (in press, 2008)
10. Tsui, F.C., et al.: Technical description of RODS: a real-time public health surveillance system. Journal of the American Medical Informatics Association 10(5), 399–408 (2003)

11. Wagner, M.M., et al.: Design of a national retail data monitor for public health surveillance. Journal of the American Medical Informatics Association 10(5), 409–418 (2003)
12. Gilchrist, M.J.: A national laboratory network for bioterrorism: evolution from a prototype network of laboratories performing routine surveillance. Military Medicine 165(7 Suppl. 2), 28–31 (2000)
13. Hutwagner, L.C., et al.: Using laboratory-based surveillance data for prevention: an algorithm for detecting Salmonella outbreaks. Emerging Infectious Diseases 3(3), 395–400 (1997)

Analysis and Prediction of Epidemiological Trend of Scarlet Fever from 1957 to 2004 in the Downtown Area of Beijing

Yanhui Shen, Chu Jiang, and Zhe Dun

Haidian Centers for Disease Prevention and Control of Beijing,
100037 Beijing, China

Abstract. Model fitting and prediction of scarlet fever in the downtown area of Beijing was conducted through time series analysis to describe the epidemiological trend. A database was built and data were fitted with Excel. ARIMA analysis and prediction were made with SPSS. Data from 1957 to 2001 were used for modeling. Data from 2002 to 2004 were used to validate the precision of the model. The incidence of scarlet fever in the downtown area of Beijing since 1957 declined, although fluctuations were apparent. There were two epidemic periods of scarlet fever, at 6.8571 years and 4.8000 years ($P<0.10$). The incidence in 2008 was predicted as 4.707/100,000 (95% confidence level: 1.379, 16.071; R^2 =0.296). Scarlet fever in Beijing is a periodical epidemic. The data of scarlet fever can be analyzed by ARIMA model.

Keywords: Scarlet fever, incidence, model.

1 Introduction

Scarlet fever is an air-born disease which can not be prevented by vaccination yet. As a class B notifiable disease, its incidence rate ranked the tenth among all notifiable diseases in China in 2004.

The epidemic trend of scarlet fever in Beijing has not been systematically studied. Will it increase or decrease? Will there be an epidemic during the 29th Olympic Games in Beijing? Has it a periodical epidemic? To answer these questions, a time series analysis using the incidence rate of scarlet fever in the downtown area of Beijing since 1957 was conducted.

2 Sources and Method

2.1 Sources

According to the "Law on notifiable infectious diseases prevention and control of the People's People's Republic of China" and "Infectious diseases report management regulation.", the Municipal Health Bureau of Beijing(MHBB) has built up a monitoring network consisting of every hospital and village clinics and community healthcare centers since 1949 to collect cases of notifiable diseases on the clinical

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 164–168, 2008.

diagnosis. The Ministry of Health categorized scarlet fever as a class "B" notifiable disease. Once a confirmed or suspected case is found in a hospital, doctor should fill in a report card instantly and within 12 hours send it to local district Center for Disease Prevention and Control (CDC) by mailing paper-based card from 1949 to 1986 or by nationwide computerized reporting network system since 1987. District CDC audited data daily and transfered the data to the municipal CDC daily and to MHBB. To guarantee the data quality, both self-inspections of hospitals and supervision of health inspection institutes were employed to prevent under-reporting and misdiagnosis. Every day CDC checks the logical errors and late-reporting. Totally the report timeliness rate is 95.69%, the under-reporting rate is below 5%, duplicate-reporting rate is 2.82%. So, the data is reliable and representative. Demographic information was provided by the Beijing Municipal Bureau of Public Security.

2.2 Diagnostic Criteria

The diagnosis of scarlet fever was based on the epidemic data (local prevalence of scarlet fever; contact history during incubation period with patients who have scarlet fever, tonsillitis, angina, tympanitis or erysipelas), symptoms and signs (fever, angina, strawberry-like tongue; rash, Pastia lines and a pale area around the lips one to two days after onset; disappearance of rash and beginning of desquamation two to five days after onset) and laboratory examinations.

2.3 Statistical Analysis

We used the EXCEL software to set up a database. We carried out Auto-Regressive Integrated Moving-Average (ARIMA) prediction and analysis with the SPSS 13.0. The data from 1957 to 2001 were used to construct a model, and the data from 2002 to 2004 were used to test the precision of the model.

Basic ideology of the model: A time series was defined as y(t), with t as time and y(t) as incidence rate of the t year. Baseline was defined as t=0, with t equals to current year minus 1957, (t=1, 2 …n, n=48).

An independent variable for time was created as follows: t2=t**2, t3= t**3, t4=t**4, t5=t**(-1), t6=t**(-2), t7=t**(-3), t8=t**(-1/2), t9=exp(-t), t10=ln(t). Y(t) was split into two parts : randomized and non-randomized parts. The randomized part was split into another two parts, which were f(t) and w(t). The formula was expressed as Y(t)=f(t)+w(t)+x(t), among which, x(t) was a randomized variable. Ln(y) was set as dependent variable and t, (t2-t10) were set as independent variables. Linear regression was conducted, and then using the SPSS backward elimination method to select variables, the main variables were identified. Using probability theory, the peak value of residual spectrum density was calculated to estimate the epidemic period. The Fisher formula was used to test the epidemic period. A periodical regression variable was created and a regression equation was established. If serials were uncorrelated, that is, the value of Durbin-Watson (DW) was between 1.5-2.5, then the formula above was used to predict y(t), with x(t) as the residual variable. If the serials {x(t)} were correlated, on condition that the serials were steady and randomized, we used the ARIMA procedure to determine the x(t) model for prediction. Furthermore, we used the formula above to predict y(t).

The ARIMA procedure could be categorized into three types: 1) auto-regressive model, 2) moving average model, 3) auto-regressive integrated moving-average model (ARIMA).

Using the ARIMA model, we predicted the incidence of a disease by identifying the model, estimating parameters, testing and predicting applications. We used the residual Box-Ljung Q statistic and significance test to inspect residual randomicity, and used AIC and SBC to examine goodness of fit. We compared the degree of fit between predicted and actual values from 2002-2004 to test the predictive precision.

3 Results

3.1 Long Term Trend

Figure 1 shows that, although the incidence of scarlet fever in Beijing has flunctuated, it has declined overall.

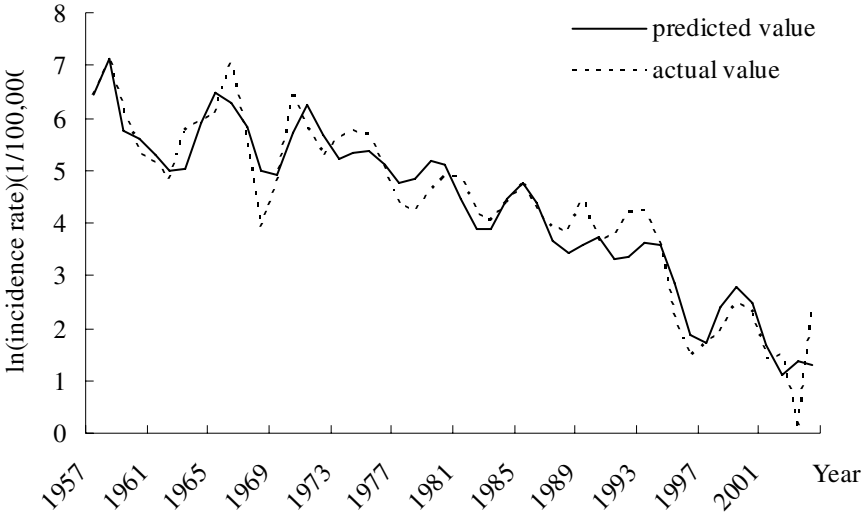


Fig. 1. Incidence rate and predicted values of scarlet fever in downtown Beijing

3.2 Identifying the Period Variable W(t)

It is especially difficult to conduct statistical tests for period because period cannot be identified by the current SPSS or SAS softwares. So, we calculated a period according to the processes introduced in *Statistics and Probability Theory*[1] *and Statistical Calculations* and *Statistical Analysis of Stationary Time Series*[2]. The steps involved first calculating the residual spectrum density then the peak value of the spectrum density curve could be used as the period. Finally, we used the Fisher formula to conduct statistical tests for each possible period. The results indicate there could exist two periods 6.8571 years and 4.8000 years ($p<0.10$).

3.3 Eliminating Period Variables and Identifying Randomized Variables

Creating period regression variables: let t11 = cos(2πt/6.8571) = cos(0.9163t), t12 = sincos(2πt/6.8571) = sin(0.9163t), t13 = cos(2πt/4.8000) = cos(1.3096t), t14 = sin (2πt/4.8000) = sin(1.3096t); ln(y) be the dependent variable, t and t2-t14 be independent variables, using backward elimination to conduct regressions. Among the results, t, t7, t8, t9, t11, t12, t13, R=0.947, D.W = 1.438 were selected.

3.4 Model Identification

Histogram showed the residuals were normally distributed, the residual trend indicated the residuals fluctuated across time, which means that the data is fit for ARIMA analysis. After trying time series one by one in SPSS ARIMA analysis, the results were p=0, d=0, q=1, indicating a MA(1) model.

3.5 Model Estimation

Table 1. SPSS stepwise regression equation and ARIMA parameter estimation

Variable	Regression Coefficient	Standard Error	t	P
MA1	-0.337	0.172	-1.96	0.057
t	-0.161	0.019	-8.561	0.000
t7	-7.396	2.931	-2.524	0.016
t8	-10.830	2.976	-3.639	0.001
t9	37.536	11.909	3.152	0.003
t11	-0.221	0.139	-1.588	0.012
t12	0.363	0.144	2.53	0.016
t13	0.350	0.129	2.714	0.010
Constant	10.815	1.138	9.503	0.000

3.6 Model Diagnosis

We employed the diagnosis of residual randomicity, inspected 16 points' Box-Ljung Q statistics, all p-values were greater than 0.05. Residual serials could be considered as white noises. R^2 of residual model was 0.296, standard error was 0.544, likelihood ratio was -34.445, AIC was 86.890, SBC was 103.731. Figure 1 shows that predicted values and actual values fit well.

3.7 Prediction

With the model described above, we predicted the incidence rate of scarlet fever in 2008 to be 4.707/100,000 (95% confidence interval: 1.379, 16.071). Using the predicted values from 2002-2004 to test the validity resulted in actual incidence rates within 95% confidence intervals.

4 Discussion

ARIMA is an important time series analysis model of Box-Jenkins Methods [3]. The time series data is a set of random variables dependent on time except for some occasional values. The dependence relationship or autocorrelation of those random variables represents the continuity of the object. Once the autocorrelation is described by a statistic model, future values can be predicted by current and past values. Based on autocorrelation analysis, ARIMA can calculate a series of autocorrelation coefficients and partial autocorrelation coefficients. It is appropriate to study the long term trend of communicable diseases using ARIMA since it takes trend, season and random effects into account at the same time and is more reliable.

Time series model is easy to be constructed and understood. Under the condition of adequate data, it possesses highly predictive precision. However, there are also some deficiencies: time series model predicts the future value by analyzing the previous incidences of scarlet fever, by only considering the historical data without considering any other influencing factors (such as the research and development of antibiotics and antibiotics abuse). In fact, the disease incidence is influenced by multiple factors. Therefore, there will occur obvious time delay between the predicting result and the actual result when the incidence rate line has sharp changes.

Studies have shown that scarlet fever may occur regularly, but there is limited consensus. Tao found scarlet fever in Shandong Province was a periodical epidemic with the period of six to eight years by Periodogram Method [4]. Using phase space technique in chaotic dynamics, Wang carried out energy spectrum analysis and chaos analysis on the monthly incidence data of scarlet fever, epidemic encephalitis, hepatitis, typhoid etc. in Benxi City from 1955 to 1996 and found scarlet fever was a periodical epidemic and non-chaotic [5]. Based on probability theory and the energy spectrum of residual error and tested with Fisher Formula, this study revealed there were two periods of scarlet fever occurrence. The major cycle is 6.9 years and the minor cycle is 4.8 years. The curve peak that was calculated by the model was almost in accord with the actual value.

References

1. Institute of Computing Technology, Chinese Academy of Science: Statistics and Probability Theory (概率统计计算). Science Press, Beijing (1979)
2. Grenander, U., Rosenblatt, M.: Statistical analysis of stationary time series. John Wiley and Sons, New York (1957)
3. Sun, Z.Q., Xu, Y.Y.: Medical Statistics (医学统计学). People's Medical Publishing House, Beijing (2006)
4. Tao, X.R.: Study on the Regularity of Scarlet Fever by Periodogram Method (用周期图法对猩红热的周期性规律的初步探讨). Journal of Preventive Medicine Information 14, 146–148 (1998)
5. Wang, Y.: Research of Correlation Fractal Dimension of Scarlet Fever Epidemic and Epidemic Cerebrospinal Meningitis (猩红热和流脑流行过程的混沌分维研究). Journal of Northeastern University (Natural Science) 20, 351–353 (1999)

Environmental Biosurveillance for Epidemic Prediction: Experience with Rift Valley Fever

Jean-Paul Chretien[1], Assaf Anyamba[2], Jennifer Small[2], Compton J. Tucker[2], Seth C. Britch[3], and Kenneth J. Linthicum[3]

[1] Walter Reed Army Institute of Research, Division of Preventive Medicine, Silver Spring, MD, USA
JeanPaul.Chretien@us.army.mil
[2] US NationalAeronautics and Space Administration-Goddard Space Flight Center, Greenbelt, MD, USA
[3] US Department of Agriculture-Agricultural Research Service, Center for Medical Agricultural, and Veterinary Entomology, Gainesville, FL, USA

Abstract. Despite established links between climate and infectious disease activity, few biosurveillance systems use climatic data to forecast epidemics. The El Niño/Southern Oscillation (ENSO) affects weather worldwide and in East Africa is associated with flooding and Rift Valley fever, a mosquito-borne viral disease of economically important livestock and humans. Following a regional ENSO-associated outbreak in 1997-1998, several agencies created a system to forecast RVF using satellite-based monitoring of ENSO and other climatic phenomena. The system generated 5 alerts since 2005. Following 3, in South Africa (2008), Sudan (2007), and East Africa (2006), RVF occurred in high-risk areas (no other RVF outbreaks were reported in monitored areas). Alerts for the Arabian Peninsula (2005) and Sudan (2005) were not followed by RVF reports, though the latter preceded a large Yellow Fever epidemic. Future directions for the system include decision analysis to guide public health interventions and extension to other climate-associated risks.

Keywords: Biosurveillance, Remote Sensing, Forecasting, Modeling.

1 Introduction

Despite many known links between climate and infectious disease activity, there are few operational biosurveillance systems that use climatic or related environmental data to forecast infectious disease epidemics [1]. Such systems could provide lead time for public health preparations, such as enhanced surveillance, risk communications, or measures to avert or lessen disease activity. Their potential benefit likely will grow, as global climate change is expected to include more frequent and severe extreme weather events [2], which may facilitate epidemics.

The El Niño/Southern Oscillation (ENSO), an irregular but natural feature of the global climate system, results from interactions between the oceans and the atmosphere across the Indo-Pacific region and affects the weather around the world. In the warm, or El Niño, phase of the cycle, sea surface temperatures are warmer than usual

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 169–174, 2008.

in the eastern-central equatorial Pacific Ocean. El Niño sometimes is followed by a cool, or La Niña, phase with colder-than-usual temperatures in the eastern-central equatorial Pacific. The warm and cool phases cycle over irregular intervals of several years but have characteristic effects on precipitation and temperature throughout much of the tropics.

El Niño is associated with increased risk of some infectious diseases [3]. For example, in East Africa, El Niño is associated with flooding and Rift Valley fever [4], a mosquito-borne viral disease that primarily affects economically important livestock, with humans infected incidentally by the mosquito vectors or by handling or consuming infected animal products. Outbreaks begin near natural depressions ("dambos") that harbor *Aedes* mosquito eggs infected directly by the parent during development. The eggs hatch with flooding, producing an initial wave of vectors; other competent vectors emerge over subsequent weeks [5] and propagate the outbreak. The largest recorded RVF outbreak, in 1997-1998, coincided with a strong El Niño. There were an estimated 89,000 human infections and hundreds of deaths in northeastern Kenya and southern Somalia [6].

Following the 1997-1998 outbreak, scientists at the US National Aeronautics and Space Administration Goddard Space Flight Center (NASA-GSFC) and the Department of Defense-Global Emerging Infections Surveillance and Response System (DOD-GEIS) initiated a partnership to forecast conditions favorable for RVF activity in Africa by monitoring ENSO and other climatic phenomena. Here, we assess the outcomes of major system alerts since 2005.

2 Methods

The RVF forecasting system uses satellite data from NASA and National Oceanographic and Atmospheric Administration (NOAA) climate and environmental observation programs to provide predictions of areas at elevated. The primary data are sea surface temperature (SST), rainfall, outgoing longwave radiation (OLR; which is correlated with cloud cover and rainfall), and Normalized Difference Vegetation Index (NDVI), which ranges from -1 to 1, with higher values indicating more dense green vegetation. NDVI is correlated with rainfall but integrates effects of other climatic parameters, responds most to sustained rather than intermittent rains, and is available globally since 1981, while ground-based rain gauge coverage is limited in Africa. SST, rainfall, OLR, and NDVI data are transformed to anomalies, or deviations from a long-term month-specific mean, to account for seasonal variability.

SST elevation in the equatorial eastern Pacific Ocean is well-known as an early El Niño indicator, and may precede heavy rainfall in East Africa by several months, while the other measures provide more proximal indications of conditions favorable for RVF. High-resolution (1 km) risk mapping is achieved using NDVI anomalies. Areas at elevated risk during month t satisfy:

$$\mathrm{NDVI}_{t-i} \geq 0.025,\ i = 0, 1, 2 \tag{1}$$

$$\frac{\sum_{i=0}^{2} NDVI_{t-i}}{3} > 0.1 \tag{2}$$

where $NDVI_t$ is the NDVI anomaly at month t. The first requirement is for 3 consecutive months with NDVI anomalies above the range of typical variation in desert areas, reflecting sustained rainfall; while the second is for an average anomaly over the 3-month window large enough to indicate heavy rainfall [7].

Updated forecasts are available monthly or more frequently if appropriate, on the DOD-GEIS website (www.geis.fhp.osd.mil/). Forecasts and alerts also are communicated to public health agencies that can act on them in at-risk areas. Important partners in responding to forecasts and alerts include the World Health Organization (WHO), Food and Agriculture Organization of the United Nations (FAO), the US Centers for Disease Control and Prevention (CDC)'s International Emerging Infections Program (IEIP) in Kenya, and two members of the DOD-GEIS network: the US Army Medical Research Unit-Kenya (USAMRU-K) in Nairobi and the US Naval Medical Research Unit-3 (NAMRU-3) in Cairo.

We defined system alerts as ones in which off-cycle (i.e. not regular monthly) warnings of RVF activity were issued to partner public health organizations. We assessed the accuracy of the predictions by searching reports from WHO and the World Organization for Animal Health corresponding to the countries where RVF activity was predicted. Successful predictions were ones in which RVF activity was reported to have occurred in an area while the area was flagged as high risk.

3 Results

RVF activity was reported in high risk areas following 3 of 5 system alerts (Table).

Table. RVF alerts and outcomes

Country or region	Outcome
South Africa (2008)	RVF in risk area (livestock, Jan-Mar 08)
Sudan (2007)	RVF in risk area (698 cases, 222 deaths, Sep 07-Jan 08)
East Africa (2006)	RVF in risk area (922 cases, 218 deaths, Dec 06-May 07)
Arabian Peninsula (2005)	No RVF
Sudan (2005)	No RVF (Yellow Fever epidemic: 565 cases, 143 deaths)

3.1 Description of a Successful Prediction: East Africa, 2006-7

In September 2006, the RVF forecasting system identified indications of an impending El Niño episode, with SSTs anomalously elevated in the central-eastern Pacific ocean (+2ºC) and the western Indian ocean (+1ºC). These conditions enhanced precipitation over these areas and the Horn of Africa through November. Rainfall increased through December, with vegetation response and conditions favorable for RVF activity in large areas of northeastern Kenya and nearby areas in Somalia and Ethiopia, as well as in southern Kenya and northern Tanzania.

The RVF forecasting system released a series of epidemic warnings based on these observations. In September 2006, it issued a global, regional-scale forecast covering late 2006-early 2007 for possible El Niño-linked outbreaks, including RVF in East Africa, to the DOD-GEIS network (these forecasts were published online in December

[8]). As rainfall increased in the Horn of Africa, the FAO Emergency Prevention System for Transboundary Animal Diseases issued an RVF alert for the Horn in November, identifying areas flagged as conducive to RVF activity [9]. WHO transmitted alerts to the countries at risk for RVF activity and called for enhanced surveillance and community awareness.

USAMRU-K, in coordination with Kenya Medical Research Institute (KEMRI) and the CDC IEIP-Kenya, deployed a field team in early December to assess high-risk areas in the Garissa district of northeastern Kenya (which was experiencing severe flooding) (the risk analysis at that time is shown in Fig. 1).

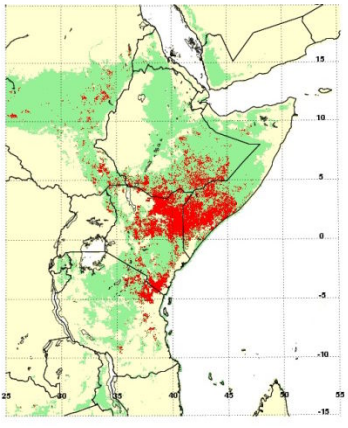


Fig. 1. RVF risk map for the Horn of Africa, December 2006. Dark shading indicates elevated RVF risk based on NDVI criteria (Eq. 1 and Eq. 2); light shading indicates areas to which the risk assessment is applied, based on ecological considerations [7].

USAMRU-K tested mosquitoes collected by the team in Garissa and from established sites in other areas, identifying RVF virus-infected mosquitoes from Garissa. The field team also investigated local reports of possible animal RVF cases and traveled with Ministry of Health staff to hospitals that recently had admitted patients with suspected RVF, obtaining specimens for testing at KEMRI.

On December 21, KEMRI confirmed RVF virus infection in specimens taken from several patients in the Garissa district [10]. The Kenya Ministry of Health initiated a response with international partners, including WHO, CDC, USAMRU-K, NAMRU-3, and the US Department of Agriculture. An intensive social mobilization campaign began in northeastern Kenya in late December, along with a locally enforced ban on animal slaughtering over most of Eastern and North Eastern Provinces (animal vaccination began in January, but by then the epidemic was waning). Frequent, high-spatial-resolution risk assessment updates were provided to facilitate targeted surveillance during the epidemic response.

Between November 30, 2006, retrospectively identified as the date of onset for the index case, and March 9, 2007, when the last case was identified, 684 cases with 155 deaths were reported in Kenya. North Eastern province, which includes the Garissa district, reported the most cases of affected provinces (N=333). Smaller RVF epidemics in Somalia and Tanzania followed the Kenya outbreaks.

3.2 Description of Two Unsuccessful Predictions

Two other RVF alerts did not precede RVF activity. Risk assessment expanded beyond East Africa to include the Arabian Peninsula following RVF outbreaks in Yemen [11] and Saudi Arabia [12] in 2000-01, the first outside Africa. These outbreaks occurred in Red Sea coastal plains that flooded following heavy rains in nearby highlands. In April 2005, positive NDVI and rainfall anomalies in the highlands exceeded those preceding the 2000-01 epidemic. The subsequent RVF alert [13] also considered the identification of RVF virus during routine sheep surveillance in western Saudi Arabia in 2004 [14], increasing the chance that facilitating environmental conditions could lead to larger-scale RVF activity. However, no RVF activity was reported following this alert.

Another unsuccessful RVF prediction, for Sudan in 2005, was followed by a large-scale Yellow Fever epidemic in areas flagged as high risk for RVF [15] (Yellow Fever, like RVF, is a mosquito-borne viral disease).

4 Discussion

The RVF forecasting system accurately predicted RVF activity in 3 of 5 alerts since 2005 (there were no other RVF outbreaks reported in areas covered by the system during this time). For the Arabian Peninsula, where RVF activity was incorrectly predicted in 2005, the risk model, which is based on experience in Kenya, may require adaptation to account for different vector ecology, human behavior, or other factors. The Yellow Fever epidemic that affected areas of Sudan flagged as high risk in 2005 may have been related to RVF-facilitating conditions, as Yellow Fever also is transmitted by mosquitoes whose populations increase following flooding.

We are assessing the costs and benefits of actions that could be taken in response to RVF predictions. While the early warning of the 2006-7 East Africa epidemic lead to enhanced mosquito surveillance and active human surveillance, a lead time of 4-6 months would, in theory, allow for large-scale livestock vaccination and mosquito abatement, which might mitigate or avert an epidemic. Decision analysis may guide consideration of these interventions.

The future of environmental biosurveillance for epidemics prediction is promising. WHO has recommended development of climate-based predictive models for cholera, malaria, and several other infectious diseases [1]. Many countries have or are developing early warning systems for natural hazards [16], which may promote epidemics. Integration of epidemic prediction with such related efforts could speed the development of epidemic prediction systems and facilitate more comprehensive risk communication to communities at risk for extreme weather events.

Disclaimer

The views expressed here are those of the authors, and are not to be construed as the official views of the US Government agencies for which the authors work.

References

1. World Health Organization. Using Climate to Predict Infectious Disease Outbreaks: A Review. Geneva (2004)
2. Intergovernmental Panel on Climate Change. IPCC Fourth Assessment Report. Working Group I Report. ch. 10 (2007)
3. Kovats, R.S., Bouma, M.J., Hajat, S., Worrall, E., Haines, A.: El Nino and Health. Lancet 362, 1481–1489 (2003)
4. Linthicum, K.J., Anyamba, A., Tucker, C.J., Kelley, P.W., Myers, M.F., Peters, C.J.: Climate and Satellite Indicators to Forecast Rift Valley Fever Epidemics in Kenya. Science 285, 397–400 (1999)
5. Linthicum, K.J., Davies, F.G., Bailey, C.L., Kairo, A.: Mosquito Species Encountered in a Flooded Grassland Dambo in Kenya. Mosquito News 44, 228–232 (1984)
6. Centers for Disease Control and Prevention. Rift Valley Fever—East Africa, 1997-1998. MMWR Morb. Mortal. Wkly. Rep. 47, 261–4 (1998)
7. Anyamba, A., Linthicum, K.J., Mahoney, R., Tucker, C.J., Kelley, P.W.: Mapping Potential Risk of Rift Valley Fever Outbreaks in African Savannas Using Vegetation Index Time Series Data. Photogrammetric Engineering & Remote Sensing 68, 137–145 (2002)
8. Anyamba, A., Chretien, J.P., Small, J., Tucker, C.J., Linthicum, K.J.: Developing Global Climate Anomalies Suggest Potential Disease Risks for 2006-2007. Int. J. Health Geogr. 5, 60 (2006)
9. Food and Agriculture Organization of the United Nations. EMPRES Watch. Possible RVF Activity in the Horn of Africa (November 2006)
10. World Health Organization. Outbreaks of Rift Valley Fever in Kenya, Somalia and United Republic of Tanzania, December 2006-April 2007. Weekly Epidemiological Record 82, 169–178 (2007)
11. Centers for Disease Control and Prevention. Outbreak of Rift Valley Fever–Yemen, August-October 2000. MMWR Morb. Mortal. Wkly. Rep. 49, 1065–1066 (2000)
12. Centers for Disease Control and Prevention. Outbreak of Rift Valley Fever–Saudi Arabia, August-October, 2000. MMWR Morb. Mortal. Wkly. Rep. 49, 905–908 (2000)
13. Anyamba, A., Chretien, J.P., Formenty, P.B., Small, J., Tucker, C.J., Malone, J.L., et al.: Rift Valley Fever Potential, Arabian Peninsula. Emerg. Infect. Dis. 12, 518–520 (2006)
14. ProMED Mail. Rift Valley fever - Saudi Arabia (Jizan) (2009): OIE (2004) (October 23, 2004), `http://www.promedmail.org`
15. World Health Organization. Epidemic and Pandemic Alert and Response. Yellow fever in Sudan (November 21, 2005), `http://www.who.int/csr/don/2005_11_21/en/index.html`
16. United Nations. Global Survey of Early Warning Systems: an Assessment of Capacities, Gaps, and Opportunities towards Building a Comprehensive Global Early Warning System for All Natural Hazards (2006)

Spatial Regression-Based Environmental Analysis in Infectious Disease Informatics

Daniel D. Zeng[1,2], Ping Yan[1], and Su Li[3]

[1] Department of Management Information Systems, the University of Arizona
{zeng,pyan}@email.arizona.edu
[2] Institute of Automation, Chinese Academy of Sciences
[3] Beijing Technology and Business University
lis@lreis.ac.cn

Abstract. Studying relationships between environmental factors and infectious diseases is an important topic in public health research. The existing studies have been focused on temporal correlations among environmental risks and infectious disease outbreaks. In this paper, we advocate the importance of spatial data analysis in infectious disease-related environmental analysis. Using data from the Beijing CDC, we have conducted spatial regression analysis to study correlation between Measles occurrences and the following environmental factors: population density and proximities to railways, roads, and water systems. We report some preliminary findings concerning significant spatial autocorrelation identified from our analysis.

Keywords: Environmental analysis, infectious disease informatics, spatial regression.

1 Introduction

The public health community has accumulated significant knowledge about how various infectious diseases emerge and spread. However, studies about the impact of environmental factors on infectious disease emergence and spreading remain sketchy. It has been well-argued that environmental factors such as temperature, humidity, proximity to water body, may have compounded impact on infectious disease transmission. But detailed models studying such kind of impact are yet to developed and evaluated due to a number of reasons such as the complex nature of the interaction between the environment and disease transmission, lack of comprehensive unbiased datasets, and lack of appropriate analysis tools or informatics environments.

The existing studies have started to focus on temporal correlations among environmental risks and infectious disease outbreaks. We argue that it is critical to add the spatial dimension in these studies. In this paper, we report a research effort aiming to analyze in a spatial-temporal context correlation between several environmental factors with Measles outbreaks in Beijing. This analysis is based on spatial regression, which has been a popular statistical tool in infectious disease informatics practice but has not yet been widely applied to study the impact of environment on infectious

D. Zeng et al. (Eds.): BioSecure 2008, LNCS 5354, pp. 175–181, 2008.

diseases. As powerful software packages such as ArcGIS and GeoDA [1] are being increasingly adopted and environmental data samples are being collected more easily with GPS-enabled devices, we expect to see increasing interest in this type of spatial and spatial-temporal analysis framework. In Section 2, we briefly survey related work on environmental health risk analysis. The spatial regression analysis using the Beijing Measles dataset is presented in Section 3. We conclude in Section 4 with a summary and a brief discussion of future research.

2 Related Work

A few previous studies have aimed to identify correlation or association between one or a set of infectious diseases and certain environmental factors. We sample below four research topics. 1) Many respiratory diseases can be significantly impacted by climate, bearing obvious seasonal characteristics [2-5]. 2) Highly intensive human movement, typically associated with dense railway and road networks, has been shown to have significant impact on disease spreading [5]. 3) With an increasing population, the chances of stable transmission cycles between infected and susceptible persons are higher [6, 7]. 4) Vegetation coverage can reflect an area's environmental conditions, such as air quality [8].

Most existing work has focused on detecting temporal correlations between environmental factors and infectious disease cases. For instance, Wavelet coherency analysis and least squares regression analysis were used to identify statistical correlations between disease occurrences and climatic indices [2, 3]. However, these methods lack the ability to identify environmental factors potentially correlated with certain infectious diseases in a spatial context.

3 Measles and Environmental Factors: A Spatial Regression Analysis

While spatial data analysis has received increasing attention in many fields, including epidemiological studies, it remains underutilized in environmental analysis in the context of infectious disease informatics. One key reason lies with the difficulty of accessing environmental data and quantifying certain environmental factors. With accelerated adoption of technologies such as GIS, GPS, and remote sensing, environmental data are becoming available in a finer geographical granularity and it is very likely that environmental analysis in the public health context will become routine and lead to real-time actionable findings.

Our reported study is focused on a spatial regression analysis. In general, spatial regression first quantifies the spatial pattern through a pre-specified neighborhood structure and then examines relations between the attributes of interest and potential explanatory variables that can account for the observed spatial pattern. Spatial autocorrelation is automatically captured by this kind of analysis. In our work, we apply the spatial lag model to study the relationship between five selected spatial-geographical environmental factors and the Measles incidence rate in Beijing. In a spatial lag model, spatial autocorrelation is modeled by a linear relation between the response vector (y) and the associated

spatially lagged vector (Wy). In particular, the model can be formulated as $y = \rho Wy + X\beta + \varepsilon$, where ε is the vector of error terms that are independent but not necessarily identically distributed. The response vector y denotes the disease incidence rates at all given regions (e.g., streets). W is a spatial weighing matrix, modeling the spatial structure for each location. There are a variety of ways of specifying these weights in this matrix. For example, a matrix with elements taking value either 0 or 1 can be used to indicate whether two locations are neighbors ($w_{ij} = 1$ if locations i and j are adjacent, and zero otherwise). Distances between locations or lengths of the shared borders can be captured as weights as well [9]. X denotes the vector of explanatory variables. In our analysis, X consists of five components:

x_1: population density at street level
x_2: proximity to railways
x_3: proximity to roads
x_4: proximity to line water systems
x_5: proximity to polygon water systems

3.1 Data Collection and Preparation

The Measles dataset covering daily case reporting of Measles in 2005 from 18 administrative districts in Beijing was made available by the Beijing Centers for Diseases Control and Prevention. Each data record contains information including patient identification, home address, and hospital visit date, among others.

The environmental data using in our study cover all 18 administrative districts in Beijing and provide information on roads, railways, line water systems, and polygon surface water systems. Most of the data were obtained from the National 1:250000 terrain databases and are in the ArcGIS Shape file format. An additional dataset used in our study provides information concerning human population density at the street level, acquired from the database of China Population by Township [10]. The geocoded and digitalized information about streets, districts, and the Beijing municipal administrative boundaries were obtained from the GIS database from the Institute of Geographic Science and Natural Resources Research. The streets defined the spatial grids used to map the disease cases in our study.

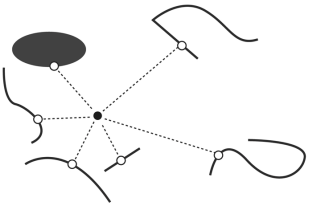
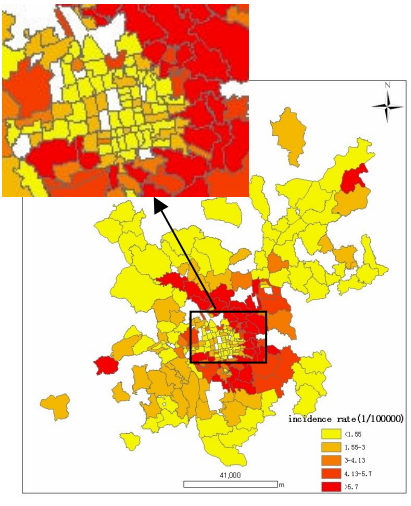
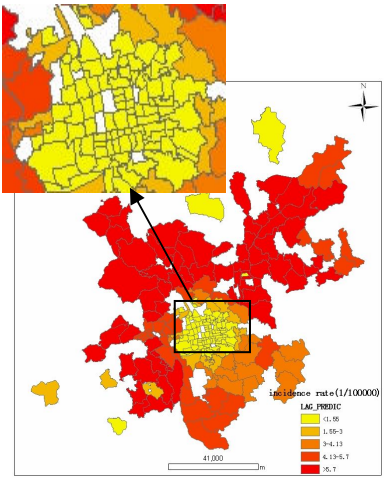


Fig. 1. Proximity from a point location to spatial objects of arbitrary shape

We hypothesize that proximity to possible sources of risks has a proportional effect on infectious disease case occurring. Values of the proximity variables are computed as the shortest distance from a location to railways, roads and water systems at the street level. ArcInfo's secondary development components, Arc Objects, provide



a) Measles incidence rates in Beijing in 2005 (above)



b) Predictions from the spatial regression model (above)

c) Predictions from the ordinary regression model (right)

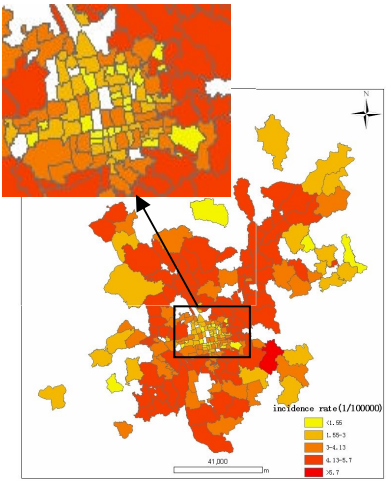


Fig. 2. Measles disease incidence rates

the corresponding programming interface to compute these shortest distances. Fig. 1 illustrates the shortest distance between a point and different spatial shapes of the spatial objects of interest. For instance, the proximity between a point and a polygon water system is the shortest distance between this point and any point on the polygon.

3.2 Results and Discussions

The resulting model generated by the spatial lag model is shown in Table 1.

Table 1. The model created by spatial lag model

	Spatial Regression
R^2	**0.483 (α=0.05)**
Result	**$y = 0.881\ W\ y + 0.668 - 80.601\ x_1 - 0.000052\ x_2 + 0.00045\ x_3 + 0.00033\ x_4 + 0.000014 x_5$** **$x_1$, W: significant**

This spatial regression achieved the R^2 measure at 0.48, indicating significant explanatory power with the spatial auto-regressive effect. Figure 2 (a) shows Year 2005 Measles incidence rates in 18 administrative districts in Beijing. The spatial grids represent administrative street divisions. Figure 2 (b) shows the predictions by the spatial regression model presented above. Figure 2 (c) shows the predictions by a regular linear regression model ($y = X\beta + \varepsilon$; results: $y = 5.94 - 114.29\ x_1 - 0.00024\ x_2 - 0.0014\ x_3 + 0.000029\ x_4 - 0.00028\ x_5$, with $R^2 = 0.067$). This regular linear regression model is included as a benchmark to evaluate the performance of the spatial regression model.

We observe that in the spatial regression model, x_1 and W are significantly correlated with Measles incidence rates. According to the model, population density x_1 is negatively correlated with Measles incidence rates. This seems counter-intuitive; however, we notice that streets with higher population density are more likely to be residential areas, typically occupied by high-rise apartments. The streets with lower population density are typically commercial centers, business offices, or parks, where intensity of people interaction can be much higher than in residential areas. The existence of significant spatial autocorrelation can be explained by the fact that people living in the neighboring streets have higher chance to interact and get infected by infected patients.

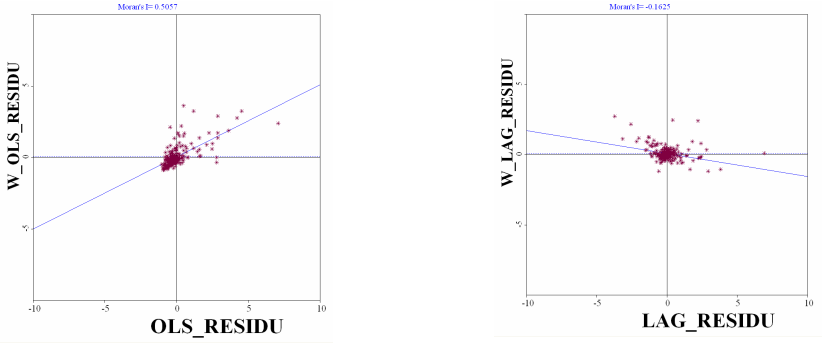


Fig. 3. Moran's I for the residuals of both the ordinary linear regression model and the spatial regression model. a) Ordinary regression: Moran's I = 0.5057; b) Spatial regression: Moran's I = -0.1625.

Moran's I is a weighted correlation coefficient used to test for global spatial auto-correlation in spatial data. When incidence rates in nearby areas are similar, Moran's I will be large and positive. When rates are dissimilar, Moran's I will be negative [11, 12]. Figure 3 indicates a strong spatial auto-correlation effect in the residual of the ordinary linear model, whereas Moran's I is significantly smaller in the residual of the spatial regression model. This indicates the effectiveness of the spatial regression model.

The existence of spatial autocorrelation might be due to frequent direct physical contacts among people who live closer. In the case of Measles, direct contact with the infected can lead to infection with high probability [13]. Indirect transmission is very rare.

4 Concluding Remarks

Effective infectious disease prevention and control requires an in-depth understanding of the disease transmission mechanisms. There is a critical need to uncover relations between environmental risk factors and disease cases, and study these relations from the point of view of data-driven research in both temporal and spatial dimensions.

This paper presents a case study of applying spatial regression analysis to analyze the relationship between Measles cases and several environmental factors using the 2005 Beijing dataset. Significant positive spatial autocorrelation and the negative impact of population density are identified. In our current research, we are analyzing datasets covering other infectious diseases and a more complete set of environmental factors, including climate, and social economic factors. For instance, in the case of Bacillary Dysentery, we have found that the following spatial factors are significant: spatial autocorrelation, proximity to railways, and proximity to polygon water sources. Our future work will also explore the predictive power of these spatial models.

Acknowledgements. The authors wish to acknowledge support from the following grants: US NSF # IIS-0428241 and # IIS-0839990; CAS #2F07C01 and #2F05N01; NNSFC #60621001, and MOST #2006CB705500.

References

1. Anselin, L., Syabri, I., Kho, Y.: GeoDa: An Introduction to Spatial Data Analysis, Spatial Analysis Laboratory, University of Illinois, Urbana-Champaign, IL (2004)
2. Subak, S.: Analysis of weather effects on variability in Lyme disease incidence in the northeastern United States. Experimental and Applied Acarology 28, 249–256 (2002)
3. Magny, G.C.d., Guégan, J.-F., Petit, M., Cazelles, B.: Regional-scale climate-variability synchrony of cholera epidemics in West Africa. BMC Infectious Diseases, 20 (2007)
4. Lillywhite, L.P.: Investigation into the environmental factors associated with the incidence of skin disease following an outbreak of Miliaria rubra at a coal mine. Occupational Medicine 42, 183–187 (1992)
5. Zhong, S.B.: Application of GIS And Remote Sensing For Study of Epidemiology of Infectious Diseases-Case Studies Of Hepatitis B And Highly Pathogenic Avian Influenza (PhD dissertation, Chinese Academy of Sciences) (2006)

6. Hoge, C., Reichler, M., Dominguez, E.: An epidemic of pneumococcal disease in an overcrowded, inadequately ventilated jail. The New England Journal of Medicine 331, 643–648 (1994)
7. Bloom, B., Murray, C.: Tuberculosis: commentary on a reemergent killer. Science 257, 1055–1064 (1992)
8. Barbour, A., Fish, D.: The biological and social phenomenon of Lyme disease. Science 260, 1610–1616 (1993)
9. Ma, C.: Spatial autoregression and related spatio-temporal models. Journal of Multivariate Analysis 88(1), 152–162 (2004)
10. China Population by Township (2002)
11. Moran, P.A.P.: Notes on continuous stochastic phenomena. Biometrika 37, 17–23 (1950)
12. Griffiths, D.A.: Spatial Autocorrelation and Spatial Filtering. Springer, New York (2003)
13. Grenfell, B.T., Bjornstad, O.N., Kappey, J.: Travelling waves and spatial hierarchies in measles epidemics. Nature 414(6865), 716–723 (2001)

Author Index