MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

INDICATOR: A Comprehensive Cyberenvironment For Infectious Disease Surveillance, Modeling, and Response

Ian S. Brooks[a], Wendy Edwards[a]

[a] *National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, USA*

Abstract

Infectious diseases continue to be a major factor in public health and emerging diseases threaten to cause a pandemic. To combat these threats it is necessary to detect outbreaks and determine the most efficient response as quickly as possible. Information technology can give health professionals the tools they need to do this, but these tools must be integrated, effective, scalable, and user friendly. NCSA is currently developing INDICATOR, an infectious disease informatics cyberenvironment, to address this need.

INDICATOR will accept data from a wide variety of sources, including hospitals, public health departments, veterinary—medicine clinics, and remote sensor networks. It will use established spatial and temporal analysis algorithms, e.g. SatScan, to combine these data sources and identify outbreaks. INDICATOR will also support epidemiological modeling with programs, e.g. EpiSims, to help develop preparedness plans and determine the most effective remediation strategies for controlling an existing outbreak. In addition it will provide collaborative, visualization, social networking, and workflow management tools. Once in production INDICATOR will provide a major new resource to the surveillance, modeling, and response communities.

Keywords:

infectious disease informatics, cyberenvironments, public health, surveillance, modeling, response

Introduction

An effective response to an outbreak of an infectious disease, whether routine or rare, naturally occurring or deliberately caused, requires the coordinated effort of healthcare providers and public health officials. To be able to respond effectively, these groups need to have rapid access to information about the outbreak including its location, size, and spread. They also need to have access to models that correlate this information with potential remediation strategies and determine how likely each strategy is to successfully contain the outbreak.

There are currently a number of software solutions available to assist in the detection of an outbreak, including the US Center for Disease Control's (CDC) National Electronic Disease Surveillance System (NEDSS) [1], Early Aberration Reporting System (EARS) [2], Laboratory Response Network (LRN) [3], BioWatch [4], and BioSense [5]; John's Hopkins' Essence [6], the Open Source Realtime Outbreak and Disease Surveillance (RODS) from Pittsburgh [7], Health Canada's Global Public Health Information Network (GPHIN) [8], and the SatScan algorithm from Harvard [9]. There are also a number of formalized surveillance systems relying on communication between physicians such as the International Society for Travel Medicine and CDC GeoSentinel group [10] and the CDC Emerging Infections Network (EIN) [11]. Most of these systems, however, work with a single type of surveillance data, such as syndromic data, mandatory reporting data, or laboratory data and focus almost exclusively on human surveillance.

There are far fewer options for planning the response to an outbreak. Developing these plans is not a simple exercise and requires conducting significant numbers of simulations based on complex modeling studies such as those being carried out by the U.S. National Institutes of Health (NIH) MIDAS (Models of Infectious Disease Agent Study) team using tools such as EpiSims [12], and EpiCast [13]. These modeling studies require the handling and processing of very large, diverse datasets and multiple complex models in a heterogeneous large scale computing environment. Ready access to high performance computers, and datasets that include biological, population, behavioral, and geographic databases requiring a significant information technology infrastructure are also necessities.

Coordinating the computing and IT resources for surveillance and modeling and integrating them into an effective system for decision makers and responders is challenging. Adding the time critical nature of a response to an outbreak, the dynamic nature of the models (where parameters need to be modified in real time), and the need to

share the results quickly and visually makes for a frontier problem that must be addressed quickly. We are working to develop a cyberenvironment to address this need.

Cyberenvironments

Today's scientists spend too much of their time moving data from place to place and format to format—time that could be better spent generating new knowledge. In the infectious disease and bioterrorism community this leads to planners working with obsolete data and modelers lacking access to real time infectious disease information from sentinel systems. What is needed is the next generation of tools to make the research process more efficient—tools that truly integrate data, models, and communities.

Cyberenvironments are this next generation. They are an integrated set of tools and services tailored to a specific discipline community that allow these communities to realize the full potential of the national cyberinfrastructure in their research, development, and teaching activities. They provide an interface to local and shared instruments and sensor networks, data stores, computational capabilities, and analysis and visualization services within a secure framework enabling management of complex projects, automation of processes, and collaboration with colleagues both near and far.

The National Center for Supercomputing Applications (NCSA) is currently working with a number of scientific and engineering communities including earthquake preparedness, atmospheric sciences, and environmental sciences to produced cyberenvironments MAEviz [14], LEAD [15], and ECID [16] respectively. These environments are aimed at revolutionizing scientific practice at the scale of communities, enabling deep collaboration within disciplines and the linking of research with application of that research. Having accurate outbreak detection algorithms and good models of the spread of infectious disease agents is of limited use if the information cannot be shared quickly and easily with all members of the wider preparedness and response community.

In these projects and others in which NCSA is involved, strong support from domain researchers, and their desire to tackle very challenging problems, is driving the development and deployment of technologies. In many of these projects there have already been clear scientific successes enabled by the new capabilities, leading to requests for additional infrastructure. Cyberenvironments involve more than the simple application of new technologies to current science and engineering practice; they enable new ways of organizing, performing, and applying research that will ultimately be shaped by the evolution of community processes.

INDICATOR

The authors together with partners from NCSA's cyberenvironments and technologies division, and external collaborators from the surveillance, modeling, and public health communities are applying the concept of cyberenvironments to infectious disease informatics to produce INDICATOR, Figure 1.

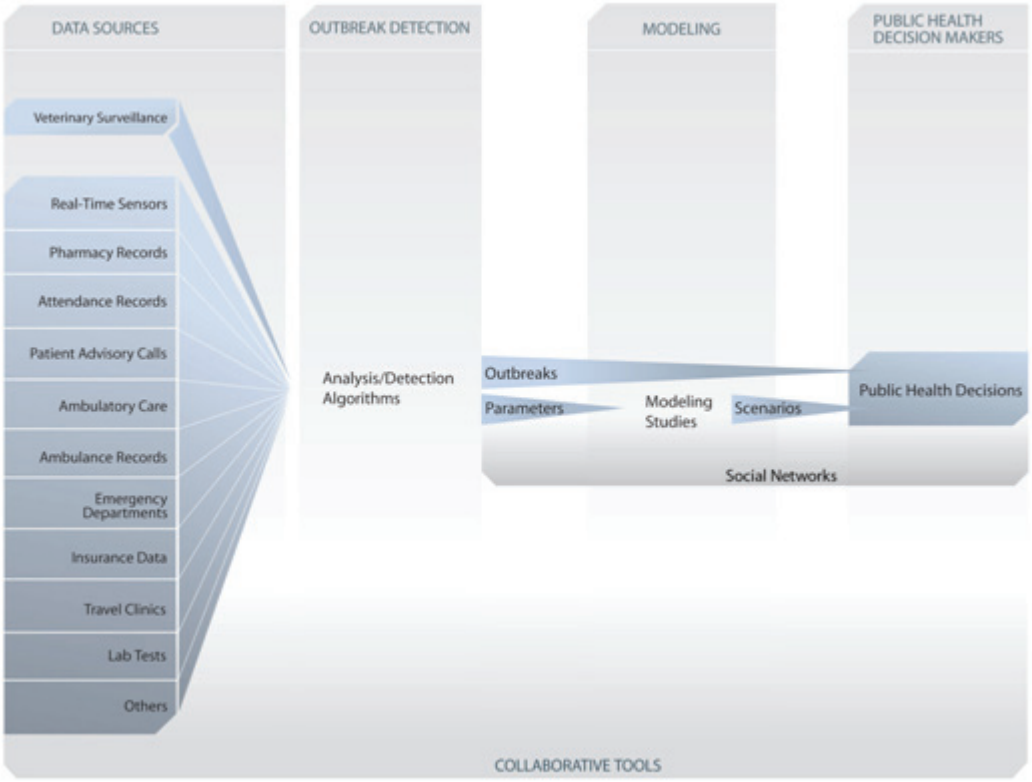

Figure 1 - Conceptual inforation flow for INDICATOR

Figure 1- Conceptual information flow for INDICATO is designed with a focus on the information needed by public health officials to make informed decisions on how best to respond to an outbreak. In most routine outbreak situations, such as the annual influenza season, it will be sufficient to detect the outbreak quickly and present the information to public health authorities in a format that they can quickly interpret. In more serious outbreaks, however, it will be essential to use parameters estimated from the outbreak detection, such as mortality and reproduction number to enable accurate modeling studies to be performed.

Effective surveillance requires the integration of multiple types of surveillance data, including syndromic, laboratory, real-time sensor networks, mandatory reporting, and veterinary data. The inclusion of veterinary data is of particular significance given the ability of wildlife monitoring to provide an early warning on the emergency of zoonotic diseases such as avian influenza and West Nile virus. Combining all of these data sources to provide a coherent picture of the health of a population is an ongoing research problem but one that INDICATOR can help algorithm developers such as Kulldorff (SatScan) address more easily. Recognizing that the data needed to identify an outbreak and model it for public health is the same data needed to advance research in surveillance INDICATOR will also provide the workflow tools needed for efficient management of algorithm testing and ensemble calculations. This will enable large numbers of computer experiments to be submitted to remote compute resources, such as NCSA supercomputers, automatically.

INDICATOR has the potential to facilitate collaboration among multiple organizations using heterogeneous technology. Users may include researchers and practitioners from many different fields, including medicine, public health, environmental science, geography, and veterinary medicine. This offers the opportunities for innovative cross-disciplinary research, but it also poses significant challenges.

Implementation

INDICATOR uses the concept of "workflow" to describe processes at multiple levels. This allows us to describe the technologies and processes in a way that can be analyzed, shared, and reused. Workflow encompasses the high-level graphical user interfaces that domain scientists use, the

low-level job execution, and the data storage and movement. Linking between views requires semantic descriptions, provenance tracking, scalable execution engines, and interfaces for managing workflows.

The workflow for an infectious disease cyberenvironment will involve heterogeneous data, including proprietary formats, processed data, and metadata. Additionally, there will be a need to support different levels of data sharing. Participants will be allowed to identify the part of their own workflows they choose to share, e.g. raw data, results, or abstract descriptions of protocols. INDICATOR will also support provenance tracking to maintain information about the origin of data.

To address these issues, we plan to develop a secure collaborative cyberenvironment and metadata/data repository, leveraging existing components created by NCSA's MAE-Viz team. Users can automatically email, upload their data, or provide an HL7 feed and INDICATOR will parse and process it as necessary. Metadata, including provenance information, will be generated and stored in the repository along with the original data. Spatial and temporal analysis software, e.g. SATSCAN [9] and WSARE [17], will be available as services.

The cyberenvironment will allow users to browse available data sets and metadata, control how their own data is shared, and select data sets and services to run. The resulting data is also stored in the repository and INDICATOR will support visual analysis, e.g., GIS maps and graphs. Users will be able to anonymize their data with FLAIM [18], a tool originally developed by NCASSR to anonymize computer security logs.

To support workflow management, we will be leveraging ECID, Environmental Cyber Infrastructure Demonstration Project, also developed at NCSA. ECID introduces the concept of "meta-workflow" which provides users with the following capabilities [16]

1. Browsing registries of data, tools, and computational resources
2. Creating meta-workflows for batch processing
3. Reuse and repurposing of meta-workflows
4. Executing meta-workflows locally or remotely
5. Incorporating heterogeneous tools and transparently linking them

For example, a user looking at emergency room chief complaint data might suspect a zoonoses-related outbreak, and want to look at veterinary data to see whether there have been unusual patterns with animal illness. The user would either browse the data or perform a metadata-based search. They could then create a new workflow using multiple services, execute it, and view the results. These results would then be stored, and the user could decide whether to share his results.

We also plan to leverage social network analysis to help users discover relevant data and services. ECID is currently exploring this technology as it develops a "dashboard" that provides a dynamic overview of the activities within the current environment [19]. This would also be useful for INDICATOR. For example, someone investigating an outbreak in one county would be able to discover recent results of a similar analysis of a neighboring county.

The LEAD (Linked Environments for Atmospheric Discovery) cyberenvironment allows the user to dynamically control real-time streaming feeds from weather sensors [15]. INDICATOR plans to leverage this technology for its remote sensing data. Weather and environmental data are often linked to infectious disease outbreaks, but the sheer volume of information has often been a problem for researchers analyzing remote sensor data. Allowing the user to dynamically choose real-time feeds would provide a solution to this problem.

Preliminary work has focused on detection of outbreaks from chief complaint data obtained from calls to a local patient advisory nurse. Having evaluated many of the available syndromic surveillance systems and algorithms SatScan and WSARE were selected for initial implementation as services based on their portability and functionality. The evolving nature of a cyberenvironment means that others may be added without disruption to the existing services and in fact INDICATOR provides an ideal platform for comparison of different algorithms.

Conclusions

Cyberenvironments have been employed successfully in multiple scientific and technical domains. These include the earthquake preparedness domain that uses data feeds to detect an earthquake, computational modeling to predict structural damage, and high end visualization to present results to emergency response authorities. The knowledge and experience gained with other cyberenvironments will be applied to the infectious disease domain to develop INDICATOR which will have analogous capabilities to use data feeds to detect an outbreak, computational modeling to predict disease spread, and high end visualization to present results to emergency response authorities. Over the next year we expect to implement workflows for SatScan and EpiSims, implement a cybercollaboratory for EIN, and move our chief complaint surveillance component into production.

Acknowledgments

The authors would like to thank Martin Kulldorff and Ken Kleinman of Harvard Medical School for their help with surveillance

algorithms and identifying their workflow needs; Steven Eubank and Madhav Marathe of Virginia Tech University for help understanding EPISims and the needs of the modeling community; David Freedman of GeoSentinel for help in understanding international surveillance issues; Phil Polgreen from EIN for helping us design the collaboratory features of INDICATOR; Lynne Reagan, Cora Musial, Lynne Barnes, Judy Kaureauf, and Napolean Knight for insights into the needs of the public health and emergency response communities; and Jim Myers, Jay Alameda, Adam Slagell, Mark Marikos, and Lorri Coey of NCSA for providing us with their technical experience on this project.

References

[1] National Electronic Disease Surveillance System Working Group. National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health and clinical medicine. J Public Health Manag Pract. 2001 Nov;7(6):43-50.

[2] Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). J Urban Health. 2003 Jun;80(2 Suppl 1):i89-96.

[3] Gilchrist MJ. A national laboratory network for bioterrorism: evolution from a prototype network of laboratories performing routine surveillance.Mil Med. 2000 Jul;165(7 Suppl 2):28-31.

[4] http://www.milnet.com/wh/DoHS/BioWatchFactSheetFINAL.pdf

[5] Loonsk JW; Centers for Disease Control and Prevention (CDC). BioSense--a national initiative for early detection and quantification of public health emergencies. Morb Mortal Wkly Rep. 2004 Sep 24;53 Suppl:53-5.

[6] Lombardo J, Burkom H, Elbert E, Magruder S, Lewis SH, Loschen W, Sari J, Sniegoski C, Wojcik R, Pavlin J. A systems overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II). J Urban Health. 2003 Jun;80(2 Suppl 1):i32-42. Review.

[7] Tsui FC, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM. Technical description of RODS: a real-time public health surveillance system. J Am Med Inform Assoc. 2003 Sep-Oct;10(5):399-408.

[8] Mykhalovskiy E, Weir L. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can J Public Health. 2006 Jan-Feb;97(1):42-4.

[9] Kulldorff M. A spatial scan statistic. Comm. Statist. Theory Methods. 1997 (26):1481-1496,

[10] Freedman DO, Kozarsky PE, Weld LH, Cetron MS. GeoSentinel: the global emerging infections sentinel network of the International Society of Travel Medicine. J Travel Med. 1999 Jun;6(2):94-8.

[11] Executive Committee of the Infectious Diseases Society of America Emerging Infections Network The emerging infections network: a new venture for the Infectious Diseases Society of America...Clin Infect Dis. 1997 Jul;25(1):34-6.

[12] Eubank S. Network based models of infectious disease spread. Jpn J Infect Dis. 2005 Dec;58(6):S9-13.

[13] http://www.lanl.gov/orgs/tt/license/software/epicast.shtml

[14] http://maeviz.cee.uiuc.edu

[15] Droegmeier KK, Chandrasekar V, Clark R, Gannon D, Graves S, Joseph E, Ramamurthy M, Wilhelmson R, Brewster K, Domenico B, Leyton T, Morris V, Murray D, Plale B, Ramachandran R, Reed D, Rushing J, Weber D, Wilson A, Xue M, Yalda S. Linked Environments for Atmospheric Discovery (LEAD): A Cyberinfrastructure for Mesoscale Meteorology Research and Education. Preprints, 20th Conf. on Interactive Info. Processing Systems for Meteorology, Oceanography, and Hydrology, Seattle, WA, Amer. Meteor. Soc.

[16] Environmental Cyber Infrastructure Demonstration Project: A Meta-Workflow Cyber-infrastructure System Designed for Environmental Observatories http://isda.ncsa.uiuc.edu/ecid/intro.html

[17] Wong WK, Moore AW, Cooper G, Wagner M. What's Strange About Recent Events. Journal of Urban Health 2003: 80: 66-75.

[18] Slagell A, Lakkaraju, K, Luo, K. FLAIM: A Multi-level Anonymization Framework for Computer and Network Logs. 20th USENIX Large Installation System Administration Conference (LISA '06), Washington, D.C., Dec., 2006

[19] Myers JD, Dunning TH. Cyberenvironments and Cyberinfrastructure: Powering Cyber-research in the 21st Century. FOMMS 2006: Foundations of Molecular Modeling and Simulation.

Address for correspondence

Ian Brooks, Ph.D.
National Center for Supercomputing Applications
1205 West Clark St.
Urbana, IL 61801
U.S.A.
ian@ncsa.uiuc.edu