T-Cube Web Interface as a tool for detecting disease outbreaks in real-time: A pilot in India and Sri Lanka

Nuwan Waidyanatha, Chamindu Sampath
LIRNEasia,
Colombo, Sri Lanka
nuwan@lirneasia.net, chamindusampath@gmail.com

Artur Dubrawski, Maheshkumar Sabhnani, Lujie Chen
Auton Lab, Carnegie Mellon University
Pittsburgh, USA
{awd, sabhnani, lujiec}@cs.cmu.edu

Ganesan M., Vincy P.
IIT Madras's Rural Technology and Business Incubator,
Chennai, India
{ganesan, vincy}@rtbi.in

Abstract – Motivated by existing gaps and inefficiencies in the paper-based manually processed disease surveillance and notification systems in India and Sri Lanka [1], the Real-Time Biosurveillance Program (RTBP) introduces technology to health departments in Tamil Nadu, India and Sri Lanka, to answer the question: "Can software programs that detect events in public health data, and mobile phones that collect health data and receive health alerts, enable effective identification and mitigation of disease outbreaks in near-real-time?" The processes involve digitizing all clinical health records and analyzing them in near real-time to detect emerging unusual patterns in data to forewarn health workers before the diseases reach epidemic states. Health records from health facilities, namely the patient disease cases, syndrome, and demographic information, are transmitted through the mHealthSurvey mobile phone application [2] and fed in to the T-Cube data structure. T-Cube Web Interface (TCWI) is a browser-based software tool that uses the T-Cube data structure for fast retrieval and display of large volume multivariate time series and spatial information. Interface allows the user to execute complex queries quickly and to run various types of statistical tests on the loaded data [3,4,5]. Detected emerging patterns of potentially epidemic events are then disseminated to health workers in the vulnerable and surrounding areas in the form of SMS, Email, and Web published alerts [6]. This paper considers utility and importance of TCWI in support of rapid detection and mitigation of bio-medical threats in developing countries.

Key words: public health, epidemiology, event detection, spatio-temporal analysis, India, Sri Lanka.

I. INTRODUCTION

The present day government resource investments in India and Sri Lanka made for disease surveillance and notification focus predominantly on data collection. Meanwhile, little or no emphasis is put on pro-active detection of adverse health events and dissemination of the corresponding findings to the relevant health workers. Passive nature of the present paper based systems is exacerbated by being limited to reporting of only about 25 notifiable diseases, while hundreds of other diseases of public importance are neglected. Moreover, substantial latencies in field data reporting and analysis are caused by its manual delivery and processing that requires between 15 to 30 days to complete [1]. The current systems do not provide the much needed real-time capability for swift detection and reporting of emerging disease threats.

Inefficiencies of such nature often prevent effective mitigation of otherwise potentially containable adverse events, such as the alarming number of over 300 deaths due to a leptospirosis outbreak in Sri Lanka during the late 2007 and early months of 2008 [8]. This disease presents symptoms similar to those of influenza, which is a quite common ailment during the rainy seasons. The scattered number of patient complaints went unnoticed until a few fatalities were reported. Given the long incubation period of leptospirosis, by the time the epidemiologist had realized the threat, it had already reached the tipping point. An unusual number of patients complaining about specific symptoms and concentrated in a particular geographic area might have signaled the epidemiologists, enabling mitigation and preventative action. Similarly, a 2009 outbreak of chikungunya in the southern parts of Tamil Nadu went largely undetected due to inefficiencies and gaps in the existing public health reporting system [9].

Our Real-Time Biosurveillance Program (RTBP) aims at mitigating those inefficiencies. It focuses on three main topics of research: workability of technology facing constraints of deployment environments (can it work given limitations of humans and infrastructure); understanding of how the newly introduced processes affect human participants (will they help the health care workers with their work); and policy implications (what would it take for the health workers and epidemiological units ready to adopt changes, business process improvements, and re-engineering).

II. RESEARCH DESIGN

Fig. 1 shows the flow diagram of the RTBP information processing cycle. Information provided by health workers is assembled and processed, and the resulting decisions are communicated back to the health workers following a sequence

of steps. The technology and new processes have been simultaneously field tested in the state of Tamil Nadu, India and in Northwestern Province, Sri Lanka with participation of health departments, health officials, and health workers.

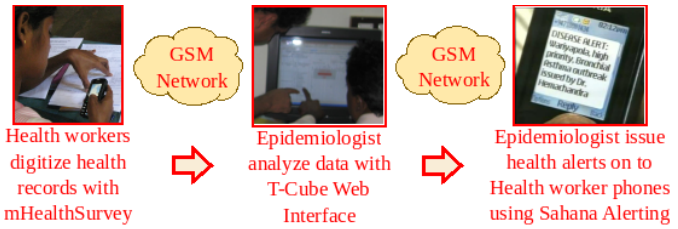

Fig 1: data digitization, analyses, and alerting work flow

III. T-CUBE TECHNOLOGY

T-Cube Web Interface (TCWI) tool is designed to efficiently visualize and manipulate large scale datasets. These types of data sets are common in epidemiology [4]. Besides executing various complex ad-hoc drill-down queries, TCWI enables a range of methods for statistical testing of multivariate temporal and spatio-temporal data. Users can manipulate and visualize data through the Time Series, Map, and Pivot Table panels.

The available statistical modeling and anomaly detection techniques include moving average, moving sum, cumulative sum, temporal scan, change scan, linear trend, peak analysis and range analysis.

The core of TCWI's Massive Screening tool is the Temporal Scan bi-variate anomaly detection algorithm [10]. It leverages efficiency of the T-Cube data structure to perform a massive number of tests of statistical hypotheses in order to find the most significantly anomalous patterns in data. The TCWI also implements Multivariate Bayesian Spatial Scan algorithm [11] that complements analyzes by testing the spatio-temporal correlations between health events. This algorithm computes the overall probability of a disease outbreak anywhere in the scope of data selected by the user separately for each day within that scope.

Fig. 2. shows the TCWI view of a respiratory tract infection (RTI) outbreak in Kurunegala District of Sri Lanka. Temporal distribution of the corresponding recorded disease cases is shown in blue. The history of the estimated probability of the RTI occurring on a given day anywhere in the District is depicted with the red line plot. Massive screening automatically identifies the periods of time of abnormally high frequency of cases of RTI, relative to cases of other diseases reported. Spatial distribution of probabilities of the RTI outbreak computed for separate villages for a few subsequent days is depicted with filled circles colored according to the value of the estimated probability. In this example, an event involving 1,408 disease cases have been detected in early March 2010. The disease propagation pattern (top of Figure 2) shows it initiating at two locations (02-Mar-2010) and spreading over to all areas (16-Mar-2010), to then subside after 27-Mar-2010.

TCWI also provides computationally efficient, interactive data summarization capability. Multidimensional data of counts of events (such as the numbers of reported disease cases) can be aggregated into a multi-way matrix view - a pivot table. Multiple attributes can be selected to denote rows and columns of the table by dragging the corresponding attribute names from the attributes list. Once a table is created and automatically filled with values, the user can click on a cell to view the corresponding time series graph, or a pie chart depicting the frequency distribution of the underlying data.

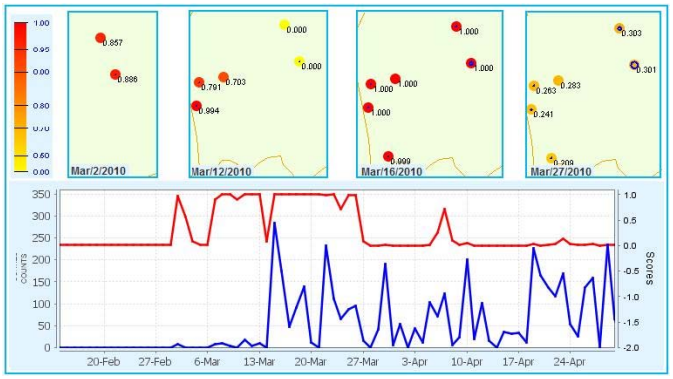

Fig 2: TCWI supports interactive visualization and statistical analysis of data enabling early detection of emerging adverse public health events.

IV. EVALUAION FRAMEWORK

To evaluate TCWI we adopted a set of generic evaluation methods known from literature [12], focusing on usability of the technology, its effects on structural or process quality, investment and operational costs, issues associated with daily operational costs, and social consequences of introduction of the technology. [13] and [14] have proposed biosurveillance system evaluation methods while [15] describes methods and key aspects of qualitative evaluation of the organizational impact of new information and communication technologies in healthcare. Methods for evaluating bioinformatics systems include subjective and objective, as well as quantitative and qualitative approaches [16].

V. DISCUSSION

Data Quality. Inconsistencies in the categorical data submitted by health workers can affect utility and performance of statistical surveillance. The incorrect entries can be to some

extent treated as statistical noise by the underlying analytic algorithms. To deal with excess of data entry errors, we defined a set of high priority diseases, symptoms, and signs. All other diseases, symptoms, and signs were labeled as "other". The shortcoming of this approach, although it does reduce the noise, is that it does not capitalize on the full analytic capabilities of TCWI to find correlations in data when some of it is clumped in a low-resolution "other" category.

In one instance, TCWI alerted of escalation of Whooping Cough in Sri Lanka. That was quite unlikely since the disease has been eradicated there for several decades. The data entry clerk had accidentally mistaken whooping cough for worm infestation. At another occasion, in India, possible dengue fever alert caught by TCWI turned out to be a result of a data entry error and it should be identified as an emerging dysentery event instead. The interactive drill-down and visualization ability of TCWI allows the users to quickly verify each alert issued by the automated detection algorithms, and to disregard those caused by the imperfections of data collection process.

Replication Study. We used publicly available Weekly Epidemiological Returns (WER) to validate our algorithms. WER data reports weekly counts of selected notifiable diseases for each of the governing districts in Sri Lanka. We used data from 2008 and 2009 and synthetically converted it to daily temporal resolution as well as simulated gender and age of patient information mimic the records collected through the RTBP, using assumed probability distributions and random drawing [17]. The data was then processed by TCWI and the results of analyzes were compared against the ground truth in WER data. We found TCWI to produce reliable signals pertaining to: (1) Leptospirosis and dengue fever events as early as two months ahead of the original time of their detection with the pre-existing processes; (2) Dengue fever emerging outside of the anticipated rainy season trend cycle; (3) Multiple instances of food poisoning not detected using the existing processes but emerging in and around the same time and area of a single originally reported incidence.

Comparison Study. Our project gathered ground truth data corresponding to the areas of live feeds of RTBP data for a span of six months. It was extracted from paper forms and registries pertaining to the existing disease reporting systems. Then, a representative set of significant events was selected and TCWI was run in retrospective mode to see how quickly it could detect them earlier. TCWI algorithms were able to detect those events within on average 1–2 days after their onset.

Over the past several months, TCWI was successfully used in the field to detect clusters of acute diarrheal disease, respiratory tract infection, dysentery, dengue, and viral fever before the health departments came to know about those incidents through the normal process.

Usability. Now, along with their routine work, health departments in India are using TCWI on an ad-hoc basis, while in Sri Lanka it is used regularly. Prior to that, during initial exposure to TCWI, the users had some difficulty in comprehending the various statistical measurement techniques offered to them by the application. With the lessons learned, through an iterative process, the RTBP team was able to customize TCWI to fit the main requirements of the health departments' objectives, while simplifying the required user activity to a small set of mouse clicks.

Future Improvements of Utility. Besides the main functionality geared for detection, TCWI has a mechanism for users to score the generated alerts. The intent is to eventually incorporate machine learning to weed out detections that are likely irrelevant, minimizing false alert rates. The TCWI alerts are manually authenticated and then ranked. Given the short span and limitation of the pilot deployment, the ranking data is inadequate for automated machine learning just yet, it however provides a qualitative measure on the utility of TCWI.

The studies of utility also involve designing procedures for scheduling TCWI event detection algorithms to execute automatically in order to periodically generate a set of alerts which then can be communicated to the appropriate recipients. These alerts will include a list of current findings, their time spans, affected locations, and probabilities or significance scores, and they will be disseminated to targeted officials via SMS and Email. These officials upon receiving the alerts could use TCWI to further verify the validity of these potentially adverse events, and to make decisions regarding any necessary responses required to mitigate the situation.

Costs and Benefits. The present day investments in public health data acquisition and analyzes in India and Sri Lanka are more than what it takes to operationalize RTBP technology. The Indian Integrated Disease Surveillance Program run by the Deputy Director of Health Services units predominantly invests resources in data collection with no emphasis given to event detection. The investments are higher in Sri Lanka because of dual entry of data at both the district and national levels. In both countries, the epidemiological statistics are used mainly for long term planning instead of ongoing surveillance. Queries involved in the kind of data visualizations and analyzes against typically large and multivariate datasets can become quite computationally costly in practice. The idle time of the Epidemiologists waiting for results is another quantifiable loss that efficient data representation techniques embedded in TCWI help eliminating. RTBP can lower the affordability and accessibility thresholds and allow for cost-effective

implementation of real-time biosurveillance capability in the two pilot locations as well as in many other developing countries. Human resources currently assigned to perform clerical work required by the paper-based data reporting and processing systems can be reassigned to monitoring and crisis response activities, contributing further to creation of an environment for swift and effective mitigation of emerging crises in public health.

VI. CONCLUSION

T-Cube Web Interface has been shown to fulfill requirements of real-time public health surveillance through extensive field testing in two developing countries. The RTBP project team intends to continue the iterative process through which the technical solutions of the user interface, technical documentation, and training regime will converge to an end state of feasible level of practical capitalization on the potential of the implemented technology. We also look forward to scaling up the coverage of the system throughout Tamil Nadu, India, and Sri Lanka, as well as to subsequently conducting similar evaluations in other countries.

ACKNOWLEDGMENTS

This work has been partially supported by the International Development Research Center of Canada (105130), US Centers for Disease Control and Prevention (R01-PH000028), the US National Science Foundation (IIS-0911032). We are grateful for enthusiastic support of the Wayamba Provincial Director of Health Services office, Sri Lanka, and of the Deputy Director of Health Service in Sivagangai District, Tamil Nadu, India, as well as of Medical Officers and Health Workers in these areas.

REFERENCES

[1] S. Prashant and N. Waidyanatha (2010). User requirements towards a biosurveillance program, Kass-Hout, T. & Zhang, X. (Eds.). Biosurveillance: Methods and Case Studies. Boca Raton, FL: Taylor & Francis, Chapter 13, pp .240-263.

[2] Gow. G., Vincy, P., and Waidyanatha, N. (2010). Using mobile phones in a real-time biosurveillance program: Lessons from the frontlines in Sri Lanka and India. 2010 IEEE International Symposium on Technology and Society (ISTAS '10), Wollongong, New South Wales, Australia.

[3] M. Sabhnani, A. Moore, and A. Dubrawski (2007). Rapid processing of ad-hoc queries against large sets of time series. Advances in Disease Surveillance, Advances in Disease Surveillance, Vol 2, 2007.

[4] S. Ray, A. Michalska, M. Sabhnani, A. Dubrawski, M. Baysek, L. Chen, J, and Ostlund (2008). T-Cube Web Interface: A Tool for Immediate Visualization, Interactive Manipulation and Analysis of Large Sets of Multivariate Time Series, AMIA Annual Symposium, 2008:1106, Washington, DC, 2008.

[5] A. Dubrawski, M. Sabhnani, S. Ray, J. Roure, and M. Baysek (2007). T-Cube as an Enabling Technology in Surveillance Applications. Advances in Disease Surveillance 4:6, 2007.

[6] G. Gow and N. Waidyanatha (2010). Using Common Alerting Protocol To Support A Real-Time Biosurveillance Program In Sri Lanka And India, Kass-Hout, T. & Zhang, X. (Eds.). Biosurveillance: Methods and Case Studies. Boca Raton, FL: Taylor & Francis, Chapter 14, pp 268-288.

[7] Ganesan M., S. Prashant, Janakiraman N., and N. Waidayanatha (2010), Real-time Biosurviellance Program: Field Experiences from Tamil Nadu, India. *IASSH conference paper*. Varanasi, Uttarpradesh, India.

[8] S. Agampodi, P. Somaratne, M. Priyantha, M. Peter (2008). An interim report of Leptospriosis outbreak in Sri Lanka – 2008, publication of the Epidedmiology Unit of Sri Lanka. Web link - http://tinyurl.com/2fpj2ct

[9] Ganesan, M., Prashant, S., Janakiraman, N., and Waidayanatha, N., (2010), Real-time Biosurviellance Program: Field Experiences from Tamil Nadu, India. ICT Academy of Tamil Nadu Journal on Communication Technology (In Press)

[10] A. Dubrawski (2009). Detection of Events in Multiple Streams of Surveillance Data. In Infectious Disease Informatics: Public Health and Biodefense, Eds. C. Castillo-Chavez, H. Chen, W. Lober, M. Thurmond, and D. Zeng. Springer-Verlag (in press).

[11] D. Neill, and G. Cooper (2009). A Multivariate Bayesian Scan Statistic for Early Event Detection and Characterization. Machine Learning (in press).

[12] E. Ammenwerth, J. Brender, P. Nykänen, H-U. Prokosch, M. Rigby, and T. Talmon (2004). Visions and strategies to improve evaluation of health information systems Reflections and lessons based on the HIS- VAL workshop in Innsbruck, International Journal of Medical Informatics, Elsevier publications, Vol. 73, pp 479-491.

[13] M. Wagner (2008). Methods for testing Biosurveillance systems, Handbook of Biosurveillance (eds. Wagner, M., Moore, M., and Aryel, R.), pp 507-515, Elsevier academic press.

[14] D. Lewis (2003) Evaluation of Public Health Informatics. Public Health Informatics and Information Systems (eds O'Carroll, P., Yasnoff, W., Ward, M., and Ripp, L), Health Informatics Series, Springer New York, pp 239-266.

[15] J. Anderson and C. Aydin (2005). Evaluating the organizational impact in healthcare information systems, Second edition, Health Informatics Series, Springer Science +Business Media.

[16] C. Friedman and J. Wyatt (2006). Evaluation methods in Bioinformatics, second edition, Health Informatics Series, Springer Science+Business Media.

[17] T. Lotze, G. Shmuli, and I. Yahav (2009). Simulating and evaluating biosurveillance datasets, Chapman and Hall (In Press).