

# Decision Support Models for Public Health Informatics

Zaruhi R. Mnatsakanyan and Joseph S. Lombardo

*With an increasing concern over emerging infectious diseases, efficient and reliable public health monitoring is critical. The prototype models described in this article were built to aid public health officials in monitoring the health of their communities by increasing situational awareness and reducing false-positive identification of disease outbreaks. This comprehensive capability is needed to bolster public health acceptance of biosurveillance systems by making the complex information environment more manageable and by achieving performance that is more robust. The models introduced in this article were built to recognize and differentiate influenza outbreaks from the other seasonal respiratory activities. The models were tested with historical data collected by the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) in the National Capital Region. Results show significant improvement in both the sensitivity and specificity of the detections compared with the ESSENCE algorithms.*

## INTRODUCTION

The increased threat of bioterrorism and naturally occurring diseases, such as pandemic influenza, continually forces public health authorities to review methods for evaluating data and reports. The objective of biosurveillance is to process large amounts of data automatically to provide the user with an awareness of the health of the community.<sup>1-3</sup> The report on bioterrorism preparedness put out by the Agency for Healthcare Research and

Quality identified several syndromic surveillance systems that are in use or in development.<sup>4</sup> The Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE), developed by APL under the sponsorship of the Defense Advanced Research Projects Agency for use in the Department of Defense Global Emerging Infections Surveillance and Response System,<sup>5</sup> is one of these systems. ESSENCE is used in both

civilian health departments and military bases throughout the United States and worldwide. In addition, ESSENCE is deployed as a network in both military and civilian installations. In the National Capital Region (NCR), the ESSENCE network allows the public health departments of Washington, DC, and several counties in the states of Maryland and Virginia to closely coordinate their work. Data collected within this network provide a unique resource for APL to work jointly with local health departments to develop new public health collaboration models that will enhance situational awareness. This requires a fundamentally new approach for the development of algorithms for biosurveillance.

Algorithms currently deployed in most of the syndromic surveillance systems provide statistical analysis of individual data streams to successfully detect data anomalies. These algorithms are capable of high sensitivity, but this sensitivity comes at the cost of excessive false positives,<sup>6</sup> especially when multiple syndrome groups and data types are processed. These algorithms, which use univariate statistical anomaly detection techniques, are good tools for quickly identifying potentially large disease outbreak events.<sup>7,8</sup> System users manage false-alarm rates by ruling out the statistical data anomalies with alternative hypotheses derived from alternative syndromic and nonsyndromic sources. Thus, the function of the detection algorithms is early detection of the potential public health events, but the user's objective is to make an informative decision based on a reliable situational-awareness picture. Another disadvantage of syndromic surveillance systems is their inability to detect an anomalous health event that presents as only a few cases because the chief complaint data are not specific enough.

The tendency to deploy multivariate algorithms<sup>9,10</sup> and the need to increase users' situational awareness mean that electronic syndromic surveillance should use decision support systems designed to directly aid clinical decision-making.<sup>11,12</sup> The increasing volume of information and its distributed nature requires the development of advanced tools to enhance users' performance in critical situations.

This article highlights three decision support models. Development of these models was initiated as an APL independent research and development project in 2007 and later funded by the Centers for Disease Control and Prevention (CDC) as one of the projects for the JHU/APL Center of Excellence in Public Health Informatics.

The first model is the Bayesian Information Fusion Model (BIFuM). Our research showed that a Bayesian Network (BN) performing multivariate information fusion can not only improve the specificity of the syndromic surveillance system but can also increase users' situational awareness by inferring data and recognizing the trends of outbreaks. The proposed method seeks to incorporate in the structure of the BN the subject

matter expert's decision-making process. The model is a probabilistic network that accepts the outputs of statistical algorithms and performs the inference to recognize disease trends in the same way that an epidemiologist would. The process includes multisource (rather than single-source) data processing and fusion. Information received from multiple health data sources is processed to detect anomalous data counts and then fused to differentiate epidemiologically significant events from mathematical anomalies in the data that are irrelevant to the epidemiologist's objective.

Furthermore, to show that emulating the collaborative decision-making process can significantly increase situational awareness, we built a second model that is a network of the BN in the decision space.

Finally, to address the problem of syndromic data quality, we introduced our third model, which integrates diagnostic information and clinical data from electronic medical records (EMRs) within a syndromic surveillance system.

All three models are decision-support rather than decision-making tools. They are not intended to replace public health experts; instead, the system analyzes available information and generates alarms that require the attention of a public health expert to decide whether additional investigation is necessary.

## BIFuM

BIFuM is a probabilistic decision support model that consists of temporal anomaly detection algorithms and the BN.<sup>13-15</sup> Temporal analysis algorithms detect statistical anomalies in the selected data queries within available data sources and generate alerts for anomalies above the threshold. The BN is used to estimate the epidemiological significance of the alerts and the public health status in a manner similar to the decision-making processes of the epidemiologist. Figure 1 shows the data-processing steps in the BIFuM. As an initial step, the model processes data counts from each data source using algorithms that ESSENCE currently deploys<sup>5</sup> to detect statistical anomalies within selected data queries. Queries were selected based on the correlations found when anomalies in the data were compared with published outbreak information from the local health departments. In addition, queries were discussed with epidemiologists to refine and verify the epidemiological relevance of selected queries to the target disease. For example, during an influenza outbreak, fever, cough, or sore throat are words commonly found in chief complaints, which suggests that one of the selected queries should be a combination of these words. After discussion with the epidemiologist, the query was refined to "fever and cough or fever and sore throat" and selected for use in the model. Queries were processed on a daily basis, and alerts were generated when an anomalous number of counts was detected.

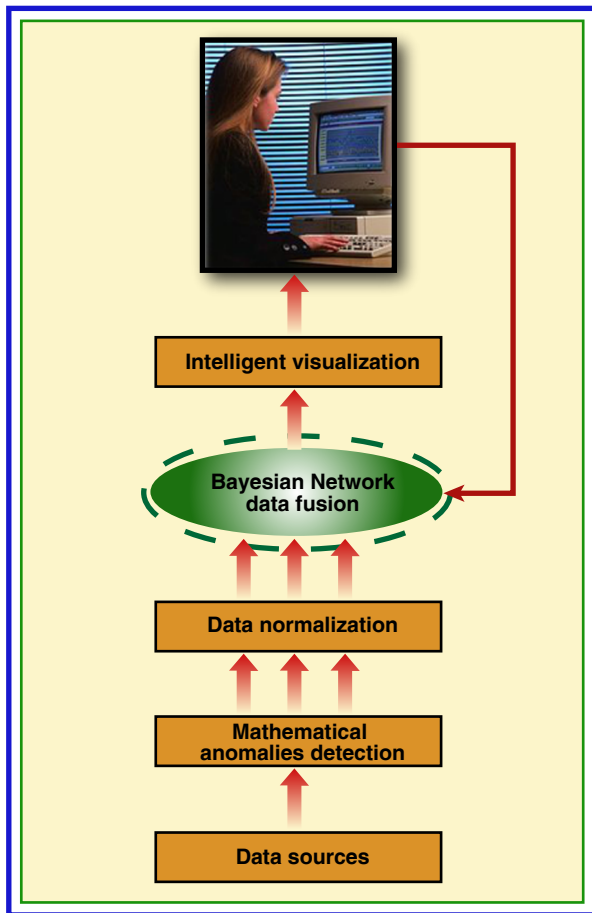


Figure 1. Data-processing diagram.

The next step is to synchronize the time offsets of detected anomalies for each query using a 7-day sliding window. The 7-day time period was chosen to incorporate a day-of-the-week effect within the model’s decision-making logic. The alert states are calculated every day for all of the selected queries. An alert state is “true” if an anomaly is detected within a query on the current day or any of the previous 6 days; otherwise, it is “false.” For example, for the query “fever” on 10 January, the alert status will be true if an anomalous number of data counts was detected on any day between 5 January and 10 January. The alert statuses are used as inputs to the BN (Fig. 2), which is the final step in data processing. The BN estimates whether detected anomalies jointly create a pattern of the epidemiologically significant event relevant to the user’s objective.

The BN is a probabilistic graphical model presented as a directed acyclic graph. Nodes of the graph represent variables, and edges represent probabilistic dependencies between variables. The BN calculates the probability of the clusters of influenza cases within infant, child, adult, and elderly age groups and the probability of the outbreak within civilian and military data sources. Figure 2 illustrates the structure of the network to calculate the conditional probability of the potential influenza-like illness (ILI) outbreak. The BN is structured so that it can classify alert states by the epidemiological significance criteria in the same way that an epidemiologist would do so manually. For example, one of the criteria is an age distribution in an ill population. If the epidemiologist

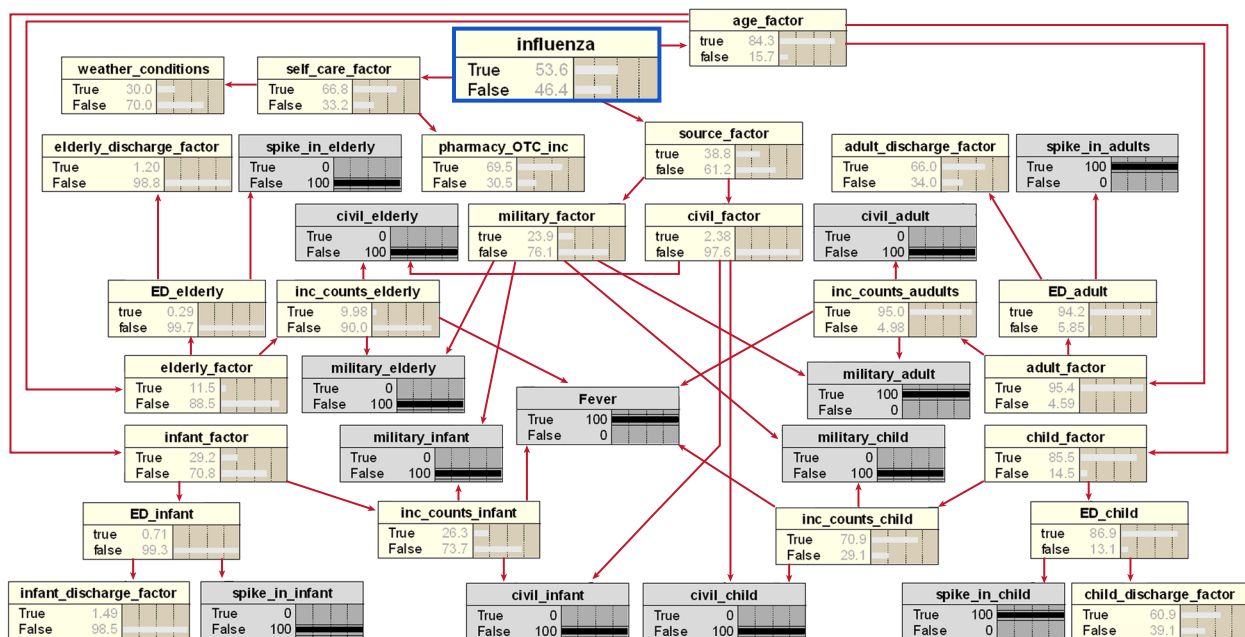


Figure 2. BN structure to calculate the conditional probability of a potential ILI outbreak. The output node is outlined in blue.

sees an anomaly in data counts for the child population, she or he may examine whether other age groups are showing an increased number of counts within the past several days, because one is unlikely to see influenza among only one age group. Our model will do the same thing—it will show a higher probability of an epidemiologically significant health event if more than one age group is affected by the disease within the past 7 days.

Another criterion is the health data source. If anomalies are detected within a particular data source (e.g., a military facility), the model will show a high probability of the outbreak based on the source factors. If node `military_factor` (Fig. 2) has high probability for the true state, this means the military population across all age groups is affected by the disease. The `civilian_factor` node shows high probability of the true state when the civilian population is affected. This information may be potentially useful in identifying a manmade outbreak that may target a specific population. For example, if `military_factor` is most likely true but `civilian_factor` is most likely false, this means the affected population is predominantly military. The sources that the model processes are either chief complaints from emergency department (ED) ILI visits or military and civilian medical office visits with an International Classification of Diseases, Ninth Edition, diagnosis code for influenza (ICD-9-487). Mathematical data anomalies within each

of the queries are detected with ESSENCE algorithms. Detections were normalized and entered into the BN, which estimated the likelihood that the detections represented an epidemiologically significant event. The results were displayed with intelligent visualization screens that provide drill-down capabilities and enable users to understand the model's behavior when a high probability of an event of interest is occurring.

The BIFuM was tested with the archived data collected over a 3-year period in the NCR. These data records were collected by ESSENCE between June 2003 and May 2006. The records include classified and processed chief complaints, discharge diagnoses and disposition from ED visits, ICD-9 diagnosis codes from visits to physicians' offices at military and civilian facilities, and over-the-counter (OTC) pharmacy transactions. All of these data are available to the users via ESSENCE.

Results showed a significant reduction in alerts compared with the ESSENCE algorithms, as well as an improvement in specificity. Figure 3 shows the number of alerts that ESSENCE generated for the query "fever and sore throat" or "fever and cough" during a 3-year period. Figure 4 shows the probabilities of influenza generated by BIFuM. The clusters when BIFuM shows a high probability of ILI correlated well with both an increase in ILI cases in data counts and the laboratory-confirmed influenza shown in Fig. 5.

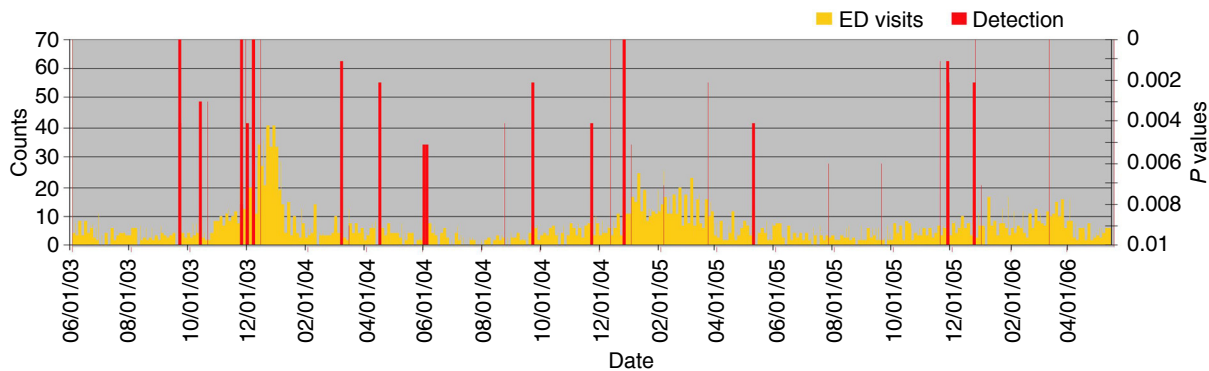


Figure 3. ESSENCE alarms and ED visits.

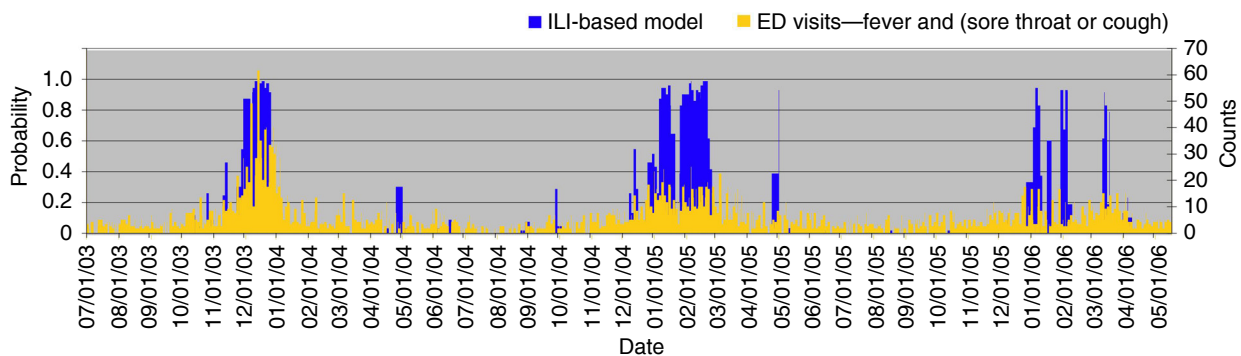
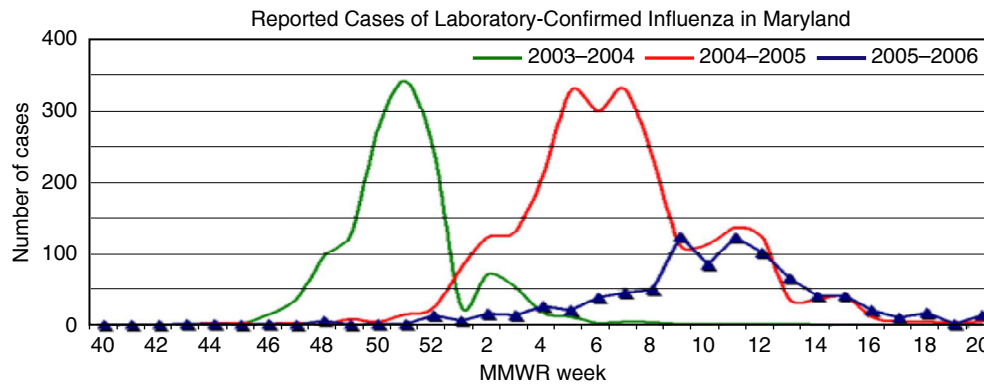


Figure 4. Probabilities of influenza outbreak.



**Figure 5.** State-reported confirmed influenza cases. MMWR, *MMWR Morbidity and Mortality Weekly Report*.

The following is a detailed description of the high probability clusters for the 3-year period:

1. For the 2003–2004 season, influenza cases first appeared during weeks 45 and 46; the system detected 46% probability of the outbreak on 12 November 2003 (week 45) and >80% probability of outbreak at the end of November (weeks 47 and 48).
2. For the 2004–2005 season, influenza cases first appeared during weeks 49 and 50, with an increase in cases during week 52 to week 1, followed by the peak at the end of January and a second wave in the beginning of February. BIFuM detected 56% probability of an outbreak on 12 December 2004 (week 50) and >80% probability of an outbreak at the beginning of January 2006 (week 1), followed by another increase in probability at the end of January. The system showed a high probability of outbreak until the end of February. However, the model showed a cluster of high probability of outbreak in the beginning of May, when there were still confirmed cases of influenza, but the number of counts was not increasing.
3. The 2005–2006 influenza season started later than usual. The first few cases appeared during week 52 and then again during week 2, with a slight increase starting during weeks 4 and 5. The peaks were on weeks 8 and 11. BIFuM showed a high probability of outbreak during week 52 (30%) and then during weeks 1, 3, and 5; the last detection was during week 10 (starting 12 March).

For all 3 years, BIFuM showed timely detection of the seasonal influenza outbreaks compared with state-reported laboratory-confirmed results. It reduced the number of epidemiologically insignificant alerts compared with the ESSENCE algorithms.

### DECISION SUPPORT NETWORK OF THE BN

Although BIFuM showed a significant increase in the performance of the syndromic surveillance system,

it used only local jurisdictions' data and did not take advantage of existing regional networks and available regional information. The Decision Support Network of the BN (DSN BN) was built to utilize available regional information and provide users with the knowledge based on the regional situational awareness picture. DSN BN was built for the NCR, which includes Washington, DC, the counties of Northern Virginia, and Montgomery County and Prince George's County in Maryland. ILI was selected as the target disease. The network consists of ILI detection models to support county-, state-, and regional-level health authorities. The DSN BN is composed of BIFuMs for each locale. In addition to the data available to the local user, the network can integrate information received from the other models representing neighboring jurisdictions in the region (Fig. 6). The situational picture in neighboring jurisdictions increases signals for the events that are happening within several counties. This is especially relevant for respiratory communicable diseases such as influenza in large metropolitan areas where very often people commute daily from one county to another. For counties that have sparse data, a timely detection of the events is generally considered to be a challenging problem. To test our regional network-centric model, we selected a county of Northern Virginia with sparse data and two neighboring counties where data were not sparse. Two fusion models for the county with the sparse data were built. The first model was receiving data from the local hospital and outpatient medical facility visits only. The second model, in addition to the local data, was receiving information from two other fusion models about the probability of influenza in the neighboring counties. Although both models detected two seasonal influenza outbreaks, the second model detected both outbreaks earlier and showed a higher probability of the influenza not only in the beginning of the outbreak but also for all of the time periods when influenza cases were increasing. This finding suggests that the model may be sensitive to the different waves of influenza.

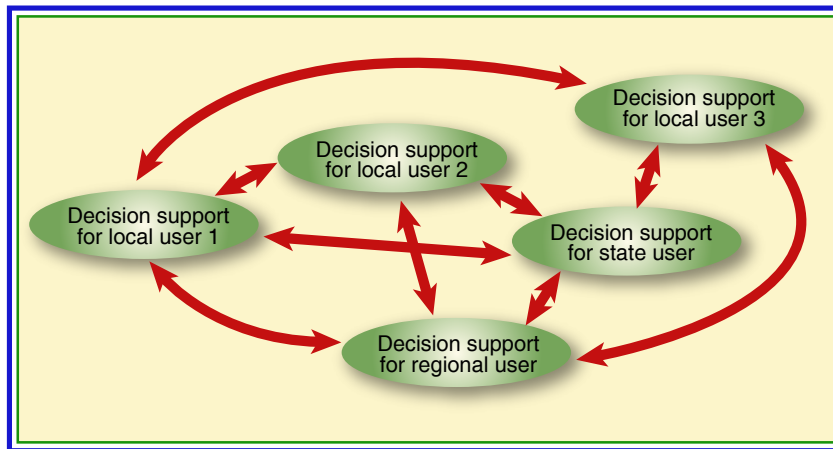


Figure 6. Network of networks.

**CLINICAL AND PUBLIC HEALTH DATA FUSION: EMR INTEGRATION IN SYNDROMIC SURVEILLANCE SYSTEM**

The data sources that traditional syndromic surveillance systems use are ED visits, chief complaints, OTC medication sales, school absenteeism, etc. These data sources and how they are organized have limitations such as data quality and lack of links between data sources. For example, there is no way to confirm that the person who buys OTC medication also had an outpatient medical facility visit. Also, available chief complaint data for hospital ED visits often do not accurately reflect the patient’s condition and present disease.

The objective of this research is to explore methods that will effectively integrate both clinical and

syndromic surveillance data and develop decision support algorithms that provide information to both public health and clinical users proactively. Multivariate information fusion algorithms based on utilizing new linked data sources from EMRs, in addition to the traditional syndromic data, will enhance outbreak recognition performance and will increase the public health user’s situational awareness. Our focus is to examine the value of laboratory, prescription medication, and radiology data linked to the patient encounter and build a model that can incorporate and utilize clinical

linked data within syndromic surveillance systems. Our main objective is to use linked clinical data from EMRs to enhance the decision support component of the syndromic surveillance system. Another objective is to increase the sensitivity of the model in order to detect initial cases of disease in the case of a bioterrorist attack.

The important new requirement to the model is to have a capability to perform data fusion with both linked and non-linked data and, based on processed individual patients’ records, derive a knowledge of the epidemiological situation in the larger population. To address this requirement, a multilayer information fusion approach was developed (Fig. 7).

Linked data have a common identifier showing that, for example, encounters, laboratory orders, and

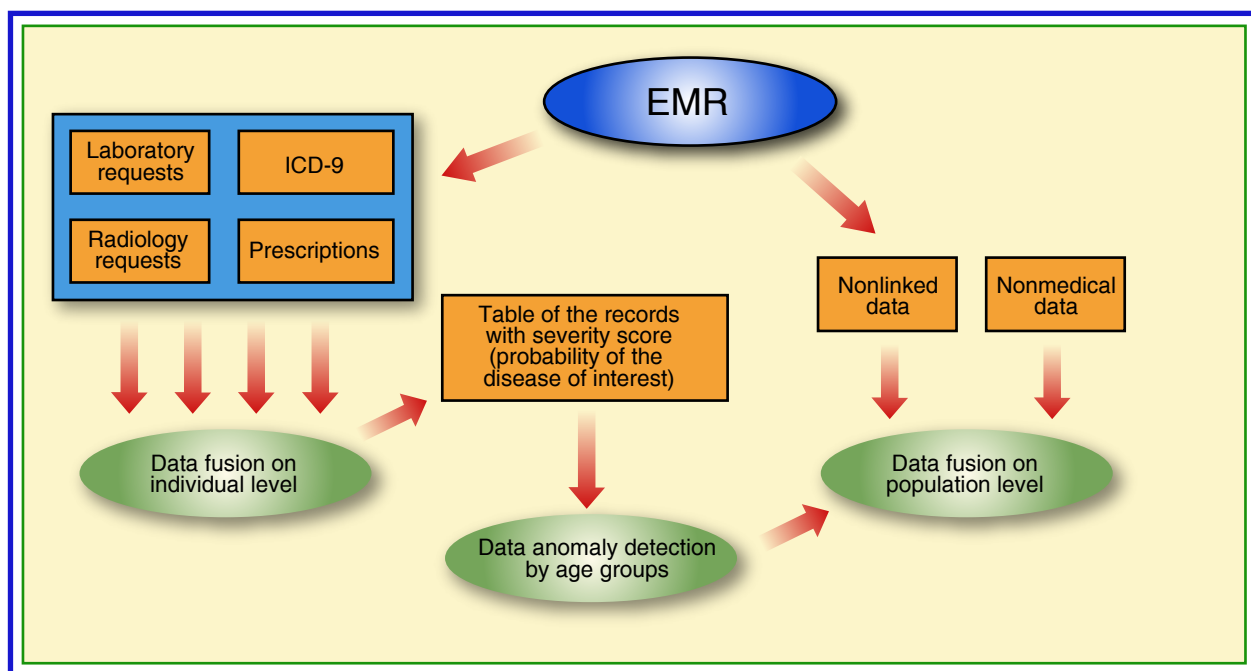


Figure 7. Clinical and public health data integration model.

radiology requests were from the same patient's record. Selected linked data sources, such as laboratory requests and results, radiology requests, and prescription medication records were associated with the individual patient encounter number. These linked data records should be processed on the individual patient level and then a severity score assigned for each patient encounter. The severity score is calculated by considering both objective and subjective factors. The objective factors are relevant ICD-9 encounters, laboratory orders, prescriptions for medication, radiology orders, a history of chronic illness, or high-risk patient status. Subjective factors are patient or physician behavior, where the model's goal is to detect unusual behavioral trends.

Severity scores are calculated for all of the patients who had an ILI encounter within the past 3-week period. Adjusted daily counts were calculated based on the severity scores and the number of daily visits. The adjusted daily count is a sum of the severity scores for the patients who have an ILI encounter. Patients are divided into infant, child, adult, and elderly age groups, and adjusted daily counts for each of the age groups are calculated. Adjusted daily counts are also calculated for patients with chronic diseases and those in the high-risk group.

Then, the model deploys temporal anomaly detection algorithms to detect anomalies in adjusted daily counts time series. Once an anomaly is detected, an alert state will be set up to true for a particular age group. For example, on 9 January, the adjusted daily count for infants was higher than the expected value, which will result in setting alert status true for infants. Alert status values for each of the age groups are input to the Population Health BN that fuses this information with unlinked data sources, such as OTC medications. The Population Health BN is ultimately the estimator of the epidemiological significance of the present situation.

## CONCLUSIONS

Our analysis showed that fusion mechanisms can significantly improve the performance of the syndromic surveillance systems. Most importantly, our models showed that the specificity of the systems can be improved without compromising sensitivity. Both specificity and sensitivity can be improved when multivariate information fusion methods and temporal anomaly detection algorithms are utilized within same model. While temporal anomaly detection algorithms can show high sensitivity on the single data streams, the information fusion algorithms can "rule out" epidemiologically irrelevant patterns, thus refining specificity of the system.

We are planning to diversify our models and create new ones that will be capable of detecting other illnesses and syndromes in addition to influenza. A new model being developed is targeting viral gastrointestinal illness (GI). Because cases of influenza are monitored and reported yearly to the public by CDC and local public health

departments, we can retrospectively evaluate the performance of each model. Although these models are showing visible improvement in the specificity of detection compared to the historical ESSENCE data, the lack of reliable ground truth for most of the syndromes does not allow us to accurately measure each model's performance. There are a number of system performance evaluation approaches<sup>16–18</sup> for syndromic surveillance systems, but they mainly focus on the early detection function of the univariate algorithms; consequently, there is a need for evaluation methodologies for multivariate syndromic surveillance systems. In addition, simulation of the outbreaks in multiple data sources remains a nontrivial problem.

Our next steps in development and evaluation of the decision support models for biosurveillance should be focused on cultivating APL's existing close working relationship with local health departments to gather feedback and reliable ground truth information to prospectively evaluate model performance and to enhance the algorithms.

The JUH/APL Center of Excellence in Public Health Informatics is addressing these fundamental issues. Our ongoing initiatives include developing a standard methodology for evaluating multivariate decision support systems for the public health domain, as well as developing simulation models to generate an outbreak based on linked clinical and public health data.

**ACKNOWLEDGMENTS:** This article was supported by Grant P01 HK000028-02 from the CDC. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CDC.

## REFERENCES

- <sup>1</sup>Lober, W. B., Karras, B. T., Wagner, M. M., Overhage, J. M., Davidson, A. J., et al., "Roundtable on Bioterrorism Detection: Information System-Based Surveillance," *J. Am. Med. Inform. Assoc.* **9**(2), 105–115 (2002).
- <sup>2</sup>Zelicoff, A. P., Brillman, J., Forslund, D. W., George, J. E., Zink, S., et al., "The Rapid Syndrome Validation Project (RSVP)," *Proc. AMIA Symp.*, 771–775 (2001).
- <sup>3</sup>Bravata, D. M., McDonald, K., Owens, D. K., Buckeridge, D., Haberland, C., et al., *Bioterrorism Preparedness and Response: Use of Information Technologies and Decision Support Systems*, Evidence Report/Technology Assessment No. 59 (Prepared by University of California San Francisco–Stanford Evidence-based Practice Center under Contract No. 290-97-0013), Agency for Healthcare Research and Quality (AHRQ), Rockville, MD, AHRQ Publication No. 02-E028 (June 2002).
- <sup>4</sup>Bartold, S. P., and Hannigan, G. G., "DXplain," *J. Med. Libr. Assoc.* **90**(2), 267–268 (2002).
- <sup>5</sup>Lombardo, J. S., Burkom, H. S., Elbert, E., Magruder, S. F., Lewis, S., et al., "A Systems Overview of the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE II)," *J. Urban Health* **80**(2), i32–i42 (Suppl. 1) (2003).
- <sup>6</sup>Marsden-Haug, N., Foster, V., Hakre, S., Anderson, A., and Pavlin, J., "Evaluation of Joint Services Installation Pilot Project and BioNet Syndromic Surveillance Systems—United States, 2004," *MMWR Morb. Mortal. Wkly. Rep.* **54**(Suppl.), 194 (2005).
- <sup>7</sup>Burkom, H. S., "Development, Adaptation, and Assessment of Alerting Algorithms for Biosurveillance," *Johns Hopkins APL Tech. Dig.* **24**(4), 335–342 (2003).

- <sup>8</sup>Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W. K., Kleinman, K., et al., "Multivariate Scan Statistics for Disease Surveillance," *Stat. Med.* **26**(8), 1824–1833 (2007).
- <sup>9</sup>Weng, K.-W., Cooper, G., Dash, D., Levander, J., Dowling, J., et al., "Use of Multiple Data Streams to Conduct Bayesian Biologic Surveillance," *MMWR Morb. Mortal. Wkly. Rep.* **54**(Suppl.), 63–69 (2005).
- <sup>10</sup>Burkom, H. S., "Biosurveillance Applying Scan Statistics with Multiple, Disparate Data Sources," *J. Urban Health* **80**(2), 157–165 (Suppl. 1) (2003).
- <sup>11</sup>Sim, I. P., Gorman, P., Greenes, R. A., Haynes, B., Kaplan, B., et al., "Clinical Decision Support for the Practice of Evidence-Based Medicine," *J. Am. Med. Inform. Assoc.* **8**(6), 527–534 (2003).
- <sup>12</sup>Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., et al., "Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review," *J. Am. Med. Assoc.* **293**(10), 1223–1238 (2005).
- <sup>13</sup>Castillo, E., Gutiérrez, J. M., and Hadi, A. S., *Expert Systems and Probabilistic Network Models*, Springer, New York (1997).
- <sup>14</sup>Pearl, J., "Fusion, Propagation, and Structuring in Belief Networks," *Artif. Intell.* **29**(3), 241–288 (1986).
- <sup>15</sup>Kahneman, D., Slovic, P., and Tversky, A. (eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, U.K. (1982).
- <sup>16</sup>Buckeridge, D. L., Burkom, H., Moore, A. W., Pavlin, J. A., Cutchis, P. N., et al., "Evaluation of Syndromic Surveillance Systems—Design of an Epidemic Simulation Model," *MMWR Morb. Mortal. Wkly. Rep.* **53**(Suppl.), 137–143 (2004).
- <sup>17</sup>Buckeridge, D. L., Switzer, P., Owens, D., Siegrist, D., Pavlin, J., et al., "An Evaluation Model for Syndromic Surveillance: Assessing the Performance of a Temporal Algorithm," *MMWR Morb. Mortal. Wkly. Rep.* **54**(Suppl.), 109–115 (2005).
- <sup>18</sup>Buehler, J. W., Hopkins, R. S., Overhage, J. M., Sosin, D. M., and Tong, V., "Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks: Recommendations from the CDC Working Group," *MMWR Recomm. Rep.* **53**(RR-5), 1–11 (2004).

# The Authors

**Zaruhi R. Mnatsakanyan** is leading research activities in the field of decision support and data fusion for The Johns Hopkins University Center of Excellence in Public Health and Disease Surveillance Program in the National Security Technology Department at APL. Dr. Mnatsakanyan has a Doctor of Engineering degree in computer science and machinery engineering from the State Engineering University of Armenia as well as master's degrees in computer science and in computer science and automated engineering from the same university. Her research interests include sensor data fusion, Bayesian Networks, pattern recognition, machine learning, and cognitive science. **Joseph S. Lombardo** has been employed by APL for the past 38 years performing research on various forms of surveillance, particularly sensors, signal coherence, background noise analysis, and data presentation. For the past 10 years, he has focused on developing and improving automated tools to enhance

disease surveillance. He has led the development of the ESSENCE disease surveillance system, which is currently being used widely by the Department of Defense, the Department of Veterans Affairs, and several state and local health departments. Mr. Lombardo has degrees in engineering from the University of Illinois at Urbana–Champaign and from The Johns Hopkins University. He was the William S. Parsons Informatics Fellow with The Johns Hopkins University School of Medicine. For further information on the work reported here, contact Dr. Mnatsakanyan. Her e-mail address is [zaruhi.mnatsakanyan@jhupl.edu](mailto:zaruhi.mnatsakanyan@jhupl.edu).



Zaruhi R.  
Mnatsakanyan



Joseph S. Lombardo