Cross-Disciplinary Consultancy to Bridge Public Health Technical Needs and Analytic Developers: Asyndromic Surveillance Use Case

Zachary Faigen[1], Lana Deyneka[1], Amy Ising[2], Daniel Neill[3], Mike Conway[4], Geoffrey Fairchild[5], Julia Gunn[6], David Swenson[7], Ian Painter[8], Lauren Johnson[9], Chris Kiley[10], Laura Streichert[9], Howard Burkom[11]

1. North Carolina Department of Health and Human Services
2. University of North Carolina at Chapel Hill, Department of Emergency Medicine
3. Carnegie Mellon University, Event and Pattern Detection Laboratory
4. University of Utah, Department of Biomedical Informatics
5. Los Alamos National Laboratory, Department of Analytics, Intelligence, and Technology
6. Boston Public Health Commission, Department of Communicable Disease Control
7. New Hampshire Department of Health and Human Services, Department of Public Health Services
8. University of Washington School of Public Health, Department of Health Services
9. International Society for Disease Surveillance
10. Defense Threat Reduction Agency, Chemical & Biological Defense Program
11. Johns Hopkins University Applied Physics Laboratory

Abstract

Introduction: We document a funded effort to bridge the gap between constrained scientific challenges of public health surveillance and methodologies from academia and industry. Component tasks are the collection of epidemiologists' use case problems, multidisciplinary consultancies to refine them, and dissemination of problem requirements and shareable datasets. We describe an initial use case and consultancy as a concrete example and challenge to developers.

Materials and Methods: Supported by the Defense Threat Reduction Agency Biosurveillance Ecosystem project, the International Society for Disease Surveillance formed an advisory group to select tractable use case problems and convene inter-disciplinary consultancies to translate analytic needs into well-defined problems and to promote development of applicable solution methods. The initial consultancy's focus was a problem originated by the North Carolina Department of Health and its NC DETECT surveillance system: Derive a method for detection of patient record clusters worthy of follow-up based on free-text chief complaints and without syndromic classification.

Results: Direct communication between public health problem owners and analytic developers was informative to both groups and constructive for the solution development process. The consultancy achieved refinement of the asyndromic detection challenge and of solution requirements. Participants summarized and evaluated solution approaches and discussed dissemination and collaboration strategies.

Practice Implications: A solution meeting the specification of the use case described above could

improve human monitoring efficiency with expedited warning of events requiring follow-up, including otherwise overlooked events with no syndromic indicators. This approach can remove obstacles to collaboration with efficient, minimal data-sharing and without costly overhead.

Keywords: asyndromic, case cluster, disease surveillance, chief complaint.

A**bbreviations**: International Society for Disease Surveillance (ISDS), Technical Conventions Committee (TCC), Defense Threat Reduction Agency (DTRA), Biosurveillance Ecosystem (BSVE), North Carolina Division of Public Health (NCDPH), Carolina Center for Health Informatics (CCHI), North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT), emergency department (ED)

Correspondence: Howard.Burkom@jhuapl.edu

DOI: 10.5210/ojphi.v7i3.6354

Copyright ©2015 the author(s)

This is an Open Access article. Authors own copyright of their articles appearing in the Online Journal of Public Health Informatics. Readers may copy articles without permission of the copyright owner(s), as long as the author and OJPHI are acknowledged in the copy and the copy is used for educational, not-for-profit purposes.

Introduction

Fifteen years into the 21st century, after worldwide publication of hundreds of articles, there is no consensus among the global disease surveillance community on preferred technical methods for public health data monitoring. Utility of such methods includes various aspects of situational awareness such as risk mapping, predictive modeling, anomaly detection, and transmission tracking. While surveillance epidemiologists frequently lack resources to address analytic needs, solution developers often lack both understanding of public health goals/constraints and the data access necessary to develop the required tools. To bridge the long-standing gap between resource-constrained scientific challenges and analytic expertise, the International Society for Disease Surveillance (ISDS) launched a Technical Conventions Committee (TCC) in January 2013 [1]. Committee activities included collection of surveillance-related use case problems from public health professionals, multidisciplinary meetings to refine these use cases, and formation of standardized requirements templates with benchmark datasets, all to facilitate the development, implementation, and publication of solution methods.

In late 2014, ISDS was awarded a contract by the Defense Threat Reduction Agency (DTRA) to enhance these activities with in-person consultancies focused on individual use cases. These activities complement the mission of the DTRA Biosurveillance Ecosystem (BSVE), an emerging capability to enable real-time biosurveillance for early warning and course-of-action analysis. The aim of BSVE is to create an unclassified virtual analyst workbench integrating health and non-health data streams and providing customized data analytics and visualization, in a cloud-based, open-source, self-sustaining web environment [2].

This paper describes the methodology of enabling inter-disciplinary and cross-agency collaboration for the advancement of public health surveillance. We provide a description of the first consultancy with its featured use case as both a concrete example and a challenge to potential developers. Previous authors have discussed strategic approaches to cross-disciplinary collaboration with workshops [3], surveys [4], and frameworks [5-7]. Many articles have been written on applicability of various statistical methods to biosurveillance [8]. However, the

authors found few articles going beyond theoretical applications to address specific needs, constraints, and operational and data limitations of health-monitoring institutions. Examples include the adaptation of the historical limits method by Levin-Rector et al. for city-level monitoring [9] and adaptation of older regression methods at the national level by Noufailly et al. [10] The recognition of the gap between the large body of analytical research and routine technical needs of public health monitors led to the formation of the TCC and the DTRA-funded initiative behind the work reported below.

The ISDS-initiated effort describes a tactical approach seeking solution methods that meet well-defined analytic needs from a work environment with known data sources and constraints. The approach is also a call to engagement for innovative, applied technologies.

Materials and Methods: Consultancy

With DTRA support, ISDS formed an advisory group of epidemiologists, technical analysts from academia and industry, and public health managers to select tractable use cases as subjects of consultancies. The consultancies' purpose is to translate use case-specific analytic needs into well-defined technical problems with shareable de-identified benchmark datasets, promote development of freely shareable solution methods, and ensure applicability in the end-user environment. Use cases considered were technical challenges posed by health departments to the TCC. These challenges were detailed in requirements templates written by health department staff including the surveillance problem description, form of output required, description of available data, and technical constraints restricting possible solutions. Advisory Group conference calls were conducted to select use cases for the consultancies based on criteria that included the public health importance and technical clarity of the proposed challenge and the likelihood of obtaining a sufficient shareable benchmark dataset. A schematic illustrating the concept of the consultancies and target use cases is shown in Figure 1.

The first selected use case was posed by the North Carolina Division of Public Health (NCDPH) and the Carolina Center for Health Informatics in the Department of Emergency Medicine at the University of North Carolina at Chapel Hill (CCHI) for use in the North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT), which receives, processes, and analyzes daily data for NCDPH. The challenge, summarized in the initial template provided in Appendix A, was to find clusters of emergency department (ED) visits of public health concern using free-text chief complaints in electronic patient records from over 100 hospitals. Problem details and the public health work environment are described below.

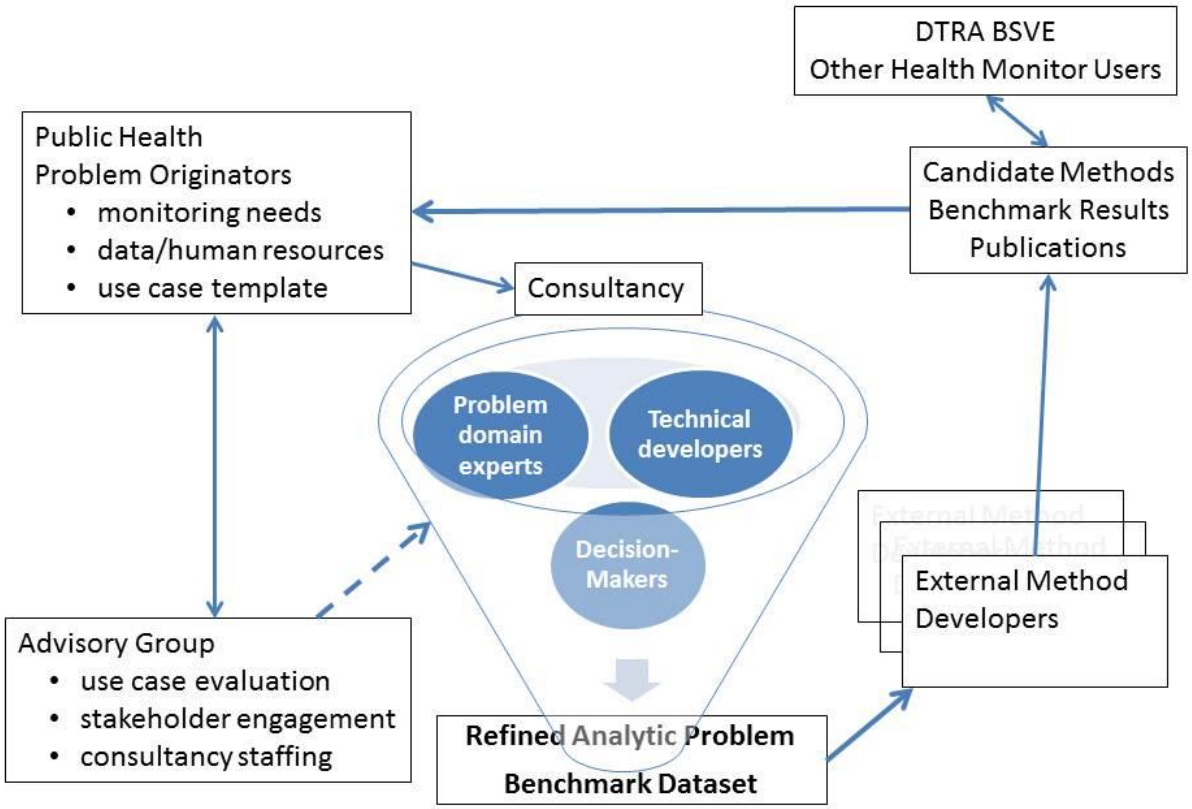

Figure 1: Conceptual diagram of inter-disciplinary consultancy to refine and disseminate target use case applications

Results: Consultancy

The consultancy was held at the UNC Gillings School of Global Public Health on June 9-10, 2015 with a planning call held on June 3. The 20 attendees included seven epidemiologists and managers from NCDPH and CCHI, three epidemiologists representing other health departments and CDC, seven analytic solution developers, and staff from ISDS and DTRA. The planning call familiarized participants with the use case, solution constraints, and the consultancy goals. Day 1 of the consultancy was devoted to the scope and details of surveillance activities at NCDPH; the functionality of the NC DETECT Web application; the use case problem of interest; the formation, composition, and exploratory analysis of the benchmark dataset; and candidate solution ideas. With this background established, Day 2 was an in-depth discussion to refine solution requirements and to propose evaluation methods. The final day also included a discussion of how to disseminate the use case problem and benchmark dataset to legitimate interested developers who would sign the NCDPH data use agreement. Tactics discussed included publicizing at conferences and on professional society forums, and staging focused workshops. The provision of adequate financial incentives for every prospective developer is not feasible, and the purpose of use case development was not to identify a single winner. However, given the number of annual publications on biosurveillance without such incentives and without

authentic datasets, expectations of a number of solvers motivated to addresses a known public health use case with data seem reasonable. See Appendix B for the consultancy agenda.

A post-consultancy survey and related discussions yielded the following lessons learned:

1. Preparatory calls should focus on details of the consultancy purpose and use case to ensure that participants come prepared and to allot more time for cross-disciplinary dialogue. Multiple respondents suggested a second, more structured preparatory call.
2. The structure of the consultancy was considered effective. Of ten responses to the post-event survey, all selected "Agree" or "Strongly Agree" to the statement, "The consultancy resulted in better definition of the use case." Nine of ten responded similarly to: "The consultancy resulted in better understanding of what is needed for a case solution", with one "Neutral" response.
3. Sufficient time should be allotted during the in-person meeting to strategize the dissemination of the use case and incentivizing solutions. Issues such as case definition and constraints therefore require substantial discussion before a consultancy. To the statement, "The consultancy ended with a clear plan to move forward on the case", seven respondents chose "Neutral", and three chose "Agree".

Materials and Methods: Use Case

Health Department Problem Environment:

North Carolina's statewide syndromic surveillance system NC DETECT provides early event detection and timely public health surveillance to authorized public health and hospital users. It was created by NCDPH in 2004 in collaboration with CCHI. A partnership of NCDPH with the North Carolina Hospital Association promoted the passage of General Statute (GS) 130A-480, which became effective January 1, 2005 [11]. This statute mandates that all NC civilian hospitals with 24/7 acute care EDs electronically report ED data elements to NCDPH at least once every 24 hours. Streaming information in NC DETECT includes near-real-time data feeds from approximately 120 North Carolina hospitals. Approximately 4.75 million ED visits, 1.3 million EMS calls, and 90,000 CPC calls are reported annually. NC DETECT uses validated syndromes for infectious diseases, injury, chronic diseases, and natural disaster response. Syndromes are defined based on ICD-9-CM final diagnosis codes and/or keywords in chief complaint and or triage notes. Aberration detection algorithms are based on CDC's Early Aberration Reporting System [12]. Signals generated by NC DETECT are analyzed, investigated, and followed up by authorized users at the state, local health department, and hospital levels. Users of NC DETECT have access to aggregate and line listing information as well as a variety of customizable reports. NC DETECT has been integral in supporting statewide disease surveillance and providing near-real-time data for population-based monitoring of illnesses and injuries before, during, and after public health emergencies.

Asyndromic Use Case Description:

NC DETECT has the flexibility to add and/or modify syndromes as needed. However, users have desired a method for detecting potentially interesting ED visit clusters *without* pre-classifying

records into syndromes. This approach would facilitate identification of clusters linked by non-symptomatic keywords indicating place names (e.g. “Midtown Café” or ”stadium”), event names (e.g., “picnic”or “football game”), or other non-medical phrases. NCDPH and CCHI developed a draft use case for this public health problem in 2014, and the requirements have been fine-tuned in subsequent meetings. In addition, a dataset was created specifically for this use case to be shared with solution developers through a Data Use Agreement with NCDPH.

Benchmark Dataset for Method Development:

The dataset includes partial records of approximately 200,000 ED visits from three hospitals in North Carolina and includes the following variables: unique visit ID, arrival date and time (altered), age group, free-text chief complaint, and non-identifying hospital code A (~31,500 records, 15.9%), B (~46,200 records, 23.3%), or C (~120,800 records, 60.8%). The number of missing values is negligible in all of the fields provided. Each chief complaint string was reviewed, and any identifiable information (names of physicians, hospitals, nursing homes, etc.) was removed. Chief complaint strings are generally short, averaging 2.7 words per record for all hospitals. The age-group classification was formed from the birth date in the original records, and these groups are shown with record frequencies in Table 1.

Table 1: Age groups with distribution of record counts in the NC DETECT benchmark dataset

Age Group	Frequency	Percent of Total
Infant (0-1 yr.)	7857	3.96
Toddler/Pre-School (2-4)	8076	4.07
Elementary School (5-9)	8023	4.04
Middle School (10-14)	6827	3.44
High School (15-18)	8517	4.29
College (19-24)	23191	11.68
Young Adult (25-44)	59380	29.91
Middle Aged (45-64)	46265	23.31
Senior (65+)	30375	15.30
Unknown	4	0.00
Total	198515	100.00

To ensure some basis for comparison of candidate methods, artificial records with related chief complaint text were added to the dataset to form injected clusters for detection. The injected cluster records were inserted in single or multiple hospitals’ data. Natural clusters are also likely to exist in the dataset, derived from authentic NC DETECT data, but verification of the number and significance of these clusters is impractical. The role of natural clusters in method evaluation will depend on the nature of these unknown clusters in submitted outputs, as discussed below.

Results: Use Case

Solution Requirements:

Solution methods must allow rapid identification of clusters of ED patient records needing public health follow-up. They may leverage patient age group and/or location in the identification of clusters as well as chief complaint and arrival date and time. The primary motivation for the use case is identification of clusters not identified through traditional record classification based on symptom-specific phrases such as "nausea", "fever", or "food poisoning". The traditional approach does not identify clusters linked by non-symptomatic keywords indicating place names (e.g., "Midtown Café" or "stadium"), event names (e.g. "picnic" or "football game"), or other event-informative non-medical phrases.

Solution methods will be run twice daily in a Windows 2012 server environment enabled with 48 GB RAM and 32 virtual processors. Methods should provide separate results for current sets of ED records from up to 121 hospitals that send approximately 16,000 records daily, and should also be applicable to records from pooled subsets of hospitals. Historical data may be used for training and adaptation; historical data are available back to 2008 for most hospitals.

Several types of flexibility are required. The requested method is intended for use by epidemiologists at the state level monitoring all hospitals and also by epidemiologists in the field monitoring subsets of 1-10 hospitals. User settings should permit adjustment of the method specificity and sensitivity from default values. The user should also be able to change the default length of the time window for inspected records, such as the previous 12, 24, 72 hours, or week. Lastly, the user should be able to influence clustering of concepts, with options to exclude terms or increase their significance. These flexibility requirements are intended to reduce the burden on human epidemiologists determining follow-up and response decisions.

The primary method output is any collection of recent ED records whose free-text information indicates some linkage potentially stimulating public health follow-up. Within NC DETECT or other surveillance systems, this output should enable rapid line listings of the included records or other visualizations such as time-series graphs or maps. Identified clusters should be ranked according to an objective significance measure for convenience, though determination of significance in practice will depend on current knowledge, concerns, and constraints of the end user.

Evaluation of Solution Methods:

Solution methods for finding asyndromic clusters will be evaluated and compared according to several criteria. Resource costs, execution time relative to the requirement of processing records from all hospitals twice a day, and ease and clarity of use will all be considered. In addition to these usability criteria, the detection performance of candidate methods will be evaluated. The benchmark dataset of 200,000 records contains a number of injected clusters of records known to the NCDPH and NC DETECT staffs. The dataset presumably contains additional authentic clusters of interest, but enumeration and verification of these is not feasible. Therefore, a twofold performance measure will be applied.

In the performance evaluation, a set of clusters produced by each method from the entire benchmark dataset will be submitted along with their significance rankings. The N clusters with the highest rankings will be evaluated, for a fixed N such as 200 for all methods. Each evaluated

cluster will be labeled by NCDPH analysts according to whether a) its records are sufficiently similar to those of an artificial cluster to call it an inject, b) it is not a result of injects but appears to merit public health follow-up, or c) it does not require follow-up. The evaluation will be blinded to the extent possible depending on the number of solutions and how soon results are available. Let n_a, n_b, and n_c be the numbers of clusters in these respective categories so that $n_a + n_b + n_c = N$, and suppose that M is the number of known injected clusters. Category c) clusters will be treated as false positives. If $n_a < M$, then the remaining injects are considered undetected by the method.

This labeling allows for two types of evaluation, on the M known injected clusters, and on an unknown number of unspecified clusters of interest. For evaluation of performance on the known injects, suppose that $r_1, r_2, \ldots, r_{n_a}$ are their ranks in descending order according to the solution method, with r_1 as the rank of the most significant inject. Then for $j = 1, \ldots, n_a$, set $f_j =$ the number of false positives ranked above r_j, so detecting the top K injects would yield f_K false positives. A plot of f_j against j is then formed for each candidate method. These plots then permit direct comparison of methods according to the known injects. In epidemiological terms, $\frac{j}{M}$ gives sensitivity, and $\frac{j}{f_j+j}$ gives positive predictive value (PPV). In the eyes of the human monitor, $\frac{1}{\text{PPV}} = \frac{f_j+j}{j}$ is the number of expected alerts for a method to produce one cluster of interest at sensitivity $\frac{j}{M}$, so a high PPV means a low alert burden.

For performance on the subjectively labeled clusters that occurred without injects, a similar procedure will be applied to the $n_b + n_c$ ranked clusters from categories b) and c). The number of true positives is unknown, and true sensitivity cannot be measured. However, the plots of f_j against j may still be used to estimate PPV as a function of the number of authentic clusters detected. Considering the authentic clusters found by each method, the NCDPH staff will weigh the detection/PPV trade-off for authentic clusters against the results from injected clusters for practical evaluation and comparison. Multiple candidate solution methods may be adopted for various purposes and for multiple types of free-text data.

Technical Approaches:

General considerations and keyword-based approaches:

Researchers with free-text analysis experience in other domains should be aware of several challenges posed by this use case. Chief complaint strings from most hospitals average 3-4 words in length. The developer may treat these strings as individual documents or pool them into blocks. For the pooled text approach, the choice of block sizes for both testing and training/learning (e.g., 1-hour, 24-hour, fixed or variable) is a key consideration. A potentially important decision is the exclusion of common or syndromic terms from the tested strings and as in solution requirements above, *ad hoc* inclusions or exclusions may be desired in practice. Lastly, solutions are to be used by epidemiologists monitoring one, several, or many facilities. In the benchmark dataset as in proposed practice, patient records have facility IDs. Thus, proposed solutions may accommodate inter-facility differences in patient schedules, variation in common free-text terms, and patient-base characteristics. Multiple solutions or parametric settings may be needed for these scenarios.

Multiple developers have considered direct, purely statistical keyword-based approaches [13-16]. A purely statistical keyword-based method with limited pre-conditioning of chief complaint text, pooling into 8-hour blocks, and a 30-day sliding baseline, showed promise for use in single facilities' data when combined with appropriate visualization [17]. For more sophisticated natural language processing or data mining strategies, added detection value should be weighed against clarity and throughput and other resource costs. The next paragraphs outline promising strategies.

Topic Models:

Another potential solution approach is based on discovering new "topics" in the free-text chief complaint data that emerge over space and time. A topic is a probability distribution over keywords, and recent topic modeling approaches such as latent Dirichlet allocation [18] enable automatic discovery of topics from text, grouping related keywords (such as nausea, vomiting, and diarrhea) into a single topic.

A recently developed "semantic scan" approach [19] incorporates spatial and subpopulation information (in the benchmark data, hospital, and age group) and can identify emerging patterns of keywords. Preliminary evaluation results on the NC DETECT dataset [20] suggest that semantic scan can identify more relevant clusters than purely keyword-based methods, since it can detect novel or unusual combinations of frequently occurring keywords as well as individual, rarely occurring keywords. These results were achieved using individual chief complaint strings as separate documents. However, topic modeling-based approaches can be computationally expensive and are sensitive to the choice of parameter values, and open questions remain as to how they can be applied most effectively in this use case context.

Feature-based clustering:

This solution type involves two stages. In Stage 1, the benchmark dataset is divided into *current* (from period that is the focus for surveillance) and *historical* (prior data for reference corpus). Surprisingly frequent words are identified in *current* data using Dunning's log-likelihood based on *historical* data [21]. Chief complaints containing these words will then be used as input for the second stage. In Stage 2, the chosen chief complaints are clustered using a feature representation based on character-based n-grams (e.g., "migraine" consists of the following 3 character n-grams: mig, igr, rai, ain, ine). This choice of feature representation is based on the published finding that under certain circumstances, character-based n-grams can better represent morphological and spelling variation than word-based methods [22]. The Stage 1 method is widely used in natural language processing and corpus linguistics [23] but can be computationally intensive when combined with the cluster analysis proposed in Stage 2.

Discussion

In addition to consultancy survey responses summarized above, free-text responses and follow-up discussions provided additional feedback. Direct communication between public health problem owners and analytic developers was informative to both groups and constructive for the solution development process. The use case originators felt that it was helpful to have a moderator from outside their health department. The participation of staff from other health departments with slightly different goals and requirements enriched the use case definition and

practical investigation of case clusters. Furthermore, potential competition among solution developer attendees did not hinder the open discussion of solution requirements and approaches.

The primary limitation of the consultancy was the brief 1.5-day time available for discussing the use case and health department environment, for refining requirements analysis, for exploring solution approaches, and for strategizing dissemination of the use case and dataset to potential developers. A limitation of the use case generation process is that funding cannot be provided to all who wish to develop solution methods, and effective incentives for solution development may vary with each use case. The only shareable dataset for the first consultancy was a large collection of ED visit records that included free-text chief complaints, age group, masked location and masked date and time. Thus, an important limitation of the asyndromic cluster detection use case is that methods that work well for finding case clusters from chief complaints may not work as well for triage notes or other data sources with longer and more complex free-text fields. Ensuring de-identification of the free text was a labor-intensive process that might be intensified for other data sources. The goal of the initial consultancy was to stimulate near-term implementation of shared methods to benefit one or a few health departments which would then inspire more generalizable development. Methods that meet the NCDPH detection requirement would need validation for application to facility data and monitoring practice in other geographic regions.

Conclusions

A solution meeting the specification of the asyndromic detection use case described above could improve human monitoring efficiency with expedited warning of events requiring follow-up, including events that would be otherwise overlooked. However, monitor expertise would remain essential for deciding on a course of action.

Attendees with health department experience discussed follow-up criteria that would be applied when evaluating a candidate cluster. For some clusters, the indicative chief complaint terms would be inconclusive, and the epidemiologist would consider additional information fields and correlations among them in search of linkages and public health significance. Such fields would include: patient age group (accounting for known quality issues such as blank age fields interpreted as 01Jan1900 and age > 115), gender, ZIP code (accounting for locations of long-term care facilities and public assistance centers), and race/ethnicity. Other clusters with phrases such as “rule out measles” or “carbon monoxide” would warrant immediate follow-up. Recent health concerns in neighboring populations or media reports would also influence perception of candidate clusters. Clustering may also identify new terms (e.g. Narcan; Ebola) indicating changing documentation and/or health care practices. These follow-up/response considerations are external to the use case challenge of initial cluster-finding. Subsequent addition of these terms to textual classifiers or triggers used for syndromic surveillance can also help keep other detection methods current.

From the above considerations, this use case exemplifies the project concept: enabling useful collaboration between methodology developer and surveillance epidemiologist without costly, long-term business arrangements, and sharing only the minimum datasets necessary for development. Anecdotal experience in this project thus far supports the assertion that combinations of data de-identification and, as necessary, truncation, perturbation, and simulation of data fields can be made feasible and acceptable for both public health problem owners and solution developers to enable such collaboration.

References

1. Coletta M, Burkom H, Johnson J, Chapman W. 2013. An ISDS-Based Initiative for Conventions for Biosurveillance Data Analysis Methods. *Online J Public Health Inform*. 5(1). http://dx.doi.org/10.5210/ojphi.v5i1.4478

2. Kiley CM, Hannan JR. The Biosurveillance Ecosystem (BSVE). 2015; http://www.dtra.mil/Portals/61/Documents/CB/BSVE%20Fact%20Sheet_04282015_PA%20Cleared.pdf. Accessed July 27, 2015.

3. Streichert LSJL. International Society for Disease Surveillance: Past Conferences. 2015; http://www.syndromic.org/annual-conference/past-isds-conferences.

4. Uscher-Pines L, Farrell CL, Cattani J, et al. 2009. A survey of usage protocols of syndromic surveillance systems by state public health departments in the United States. *J Public Health Manag Pract*. 15(5), 432-38. PubMed http://dx.doi.org/10.1097/PHH.0b013e3181a5d36b

5. Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC Working Group. *MMWR Recomm Rep.* Vol 53. United States2004:1-11.

6. German RR, Lee LM, Horan JM, Milstein RL, Pertowski CA, Waller MN. Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. *MMWR Recomm Rep.* 2001;50(RR-13):1-35; quiz CE31-37.

7. Meynard J-B, Chaudet H, Green AD, et al. 2008. Proposal of a framework for evaluating military surveillance systems for early detection of outbreaks on duty areas. *BMC Public Health*. 8(1), 146. PubMed http://dx.doi.org/10.1186/1471-2458-8-146

8. Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. 2012. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *J R Stat Soc Ser A Stat Soc*. 175, 49-82. http://dx.doi.org/10.1111/j.1467-985X.2011.00714.x

9. Levin-Rector A, Wilson EL, Fine AD, Greene SK. 2015. Refining historical limits method to improve disease cluster detection, New York City, New York, USA. *Emerg Infect Dis*. 21(2), 265-72. PubMed http://dx.doi.org/10.3201/eid2102.140098

10. Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, et al. 2013. An improved algorithm for outbreak detection in multiple surveillance systems. *Stat Med*. 32(7), 1206-22. PubMed http://dx.doi.org/10.1002/sim.5595

11. North Carolina General Assembly GS § 130A-480. Emergency department data reporting. 2015; http://www.ncga.state.nc.us/enactedlegislation/statutes/html/bysection/chapter_130a/gs_130a-480.html. Accessed 7/20/2015.

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

13. Li M, Loschen W, Deyneka L, Burkom H, Ising A, Waller A. Time of Arrival Analysis in NC DETECT to Find Clusters of Interest from Unclassified Patient Visit Records. *5*. 2013.

14. Lall R, Levin-Rector A, Mathes R, Weiss D. 2014. Detecting Unanticipated Increases in Emergency Department Chief Complaint Keywords. *Online J Public Health Inform*. 6(1). http://dx.doi.org/10.5210/ojphi.v6i1.5069

15. Taylor SA, Kite-Powell A. 2014. A Dictionary-based Method for Detecting Anomalous Chief Complaint Text in Individual Records. *Online J Public Health Inform*. 6(1). http://dx.doi.org/10.5210/ojphi.v6i1.5012

16. Walsh A, Hamby T, John TLS. 2014. Identifying Clusters of Rare and Novel Words in Emergency Department Chief Complaints. *Online J Public Health Inform*. 6(1). http://dx.doi.org/10.5210/ojphi.v6i1.5111

17. Burkom H, Elbert Y, Piatko C, Fink C. 2015. A Term-based Approach to Asyndromic Determination of Significant Case Clusters. *Online J Public Health Inform*. 7(1).

18. Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. *J Mach Learn Res*. 3, 993-1022.

19. Liu Y, Neill DB. Detecting previously unseen outbreaks with novel symptom patterns. *Emerging Health Threats J*. 2011;4.

20. Nobles M, Deyneka L, Ising A, Neill DB. Identifying Emerging Novel Outbreaks In Textual Emergency Department Data. *7*. 2015.

21. Dunning T. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput Linguist*. 19(1), 61-74.

22. Miao Y, Kešelj V, Milios E. Document clustering using character N-grams: a comparative evaluation with term-based and word-based clustering. Paper presented at: Proceedings of the 14th ACM international conference on Information and knowledge management; 10/31/2005, 2005.

23. Jingjing Liu AL, Stephanie Seneff. Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews: Focus on Statin Drugs. 2015.