Synthesizing the American Health Information Community's Minimum Data Set

Joseph Lombardo[1], John Copeland[2], Protagoras Cutchis[1]

[1]Johns Hopkins University Applied Physics Laboratory

[2]National Centers for Disease Control and Prevention

OBJECTIVE

The objectives of this presentation are to describe the need for synthetic data containing the elements of the American Health Information Community's (AHIC) Minimim Data Set (MDS). Approaches for creating synthetic data with MDS data elements will be presented and methods for insuring maintenance of confidentiality will be discussed.

BACKGROUND

One of the challenges facing developers and users of automated disease surveillance systems is being able to accurately evaluate the performance of their systems for the wide variety of public health threats that are possible. A variety of methods have been used in the past to create data sets for use in testing algorithm performance. Synthetic data has been created using agent-based simulations where data is created based on the hypothesized activity of individuals with contagious diseases [1]. This data is only as accurate as the social models and variety of assumptions which must be made permit. Real data containing elevated levels of respiratory and gastrointestinal activity have been used to evaluate the ability of algorithms to detect the elevated levels [2]. Routine unvalidated outbreaks are typically not public health emergencies and may not represent signals of interest. Another approach is to use real background data and inject a variety of different types of synthetic cases representing various types of outbreaks on top of that background [3],[4].

With the introduction of the AHIC Minimum Data Set, the public health surveillance community should have the potential to obtain greater specificity for alerts generated in automated systems. The introduction of these additional data elements increases the complexity of algorithms using linked data elements. Creating synthetic data sets that accurately estimate relationships among chief complaint, pharmacy, laboratory and radiology is an added complexity in creating synthetic outbreaks for performance evaluation.

METHODS

Some biosurveillance systems have been collecting data similar to the AHIC MDS from hospitals around the country. These data provide an excellent opportunity to study the normal background levels of disease as well as the relationships among the elements such as chief complaint, microbiology laboratory requests and results, radiology requests and results and pharmacy orders. The knowledge gained will permit creation of realistic background data streams. The data could also be modified to retain the characteristics of background while destroying all identifying features. Additional synthetic outbreak cases may be inserted into the real but modified background AHIC MDS data. Synthetic datasets would have to be reviewed by an independent statistician to certify that confidentiality of individual patients, facilities, and various levels of government (e.g, counties, states) would be maintained if the datasets were to be publicly released.

RESULTS

Burkom and Hutwagner [5] have demonstrated that the Sartwell lognormal distribution can provide a good estimate for the shape of an epidemic curve, but the curve represents the total number of cases and not the signals which could be present within each of the AHIC MDS data elements. Additional analysis is needed model the distribution among the data elements for various public health risks.

CONCLUSIONS

Synthetic AHIC MDS data injected with signals estimated from the Sartwell lognormal distribution are being investigated for creating test data sets containing elements of the AHIC MDS. These data sets are urgently needed to develop analytical tools and upgrades for future and existing automated processes to support both situational awareness and early event detection. Once acceptable background data sets are assembled representing different seasons and regions, many different signals can be added to evaluate the performance of individual components or entire systems from end to end.

REFERENCES

[1] Eubanks S, Smith J., Scalable, Efficient Epidemiological Simulation. Proc. Symposium on Applied Computing, Madrid 2002.

[2] Siegrist D, Pavlin J, Bio-ALIRT Biosurveillance Detection Algorithm Evaluation. MMWR 2004;53(Suppl);152-158.

[3] Buckeridge D, et.al., Evaluation of Syndromic Surveillance Systems: Design of an Epidemic Simulation Model MMWR 2004; 53(Suppl);137-143.

[4] Lombardo J, Buckeridge D, Thompson M, Disease Surveillance: A Public Health Informatics Approach, Ch10 Evaluating Automated Surveillance Systems, ISBN 978-0-470-06812-0, John Wiley and Sons, April 2007.

[5] Burkom H, Hutwagner L, Using Point Source Epidemic Curves to Evaluate Alerting Algorithms for Biosurveillance, Proceedings of the 2004 American Statistical Association, Statistics in Government Section , Toronto, Canada: Jan. 2005 1462-1469.

For Information: Joe Lombardo, Joe.Lombardo@jhuapl.edu