The Development of Virtual Data for Syndromic Surveillance Exercises

**Jacqueline S. Coberly, PhD, Sean P. Murphy, Steven M. Babin, MD, PhD,
Howard Burkom, PhD, Brian Feighner, MD, Jeffrey Lin, PhD, MPH, Joseph Lombardo**
The Johns Hopkins University Applied Physics Laboratory

OBJECTIVE

This paper describes a flexible modeling and simulation process that can create realistic, virtual syndromic data for exercising electronic biosurveillance systems.

BACKGROUND

On 27 April 2005, a simulated bioterrorist event—the aerosolized release of *Francisella tularensis* in the men's room of luxury box seats at a sports stadium—was used to exercise the disease surveillance capability of the National Capital Region (NCR). The objective of this exercise was to permit all of the health departments in the NCR to exercise inter-jurisdictional epidemiological investigations using an advanced disease surveillance system. Actual system data could not be used for the exercise as it both is proprietary and contains protected, though de-identified, health information about real people; nor is there much historical data describing how such an outbreak would manifest itself in normal syndromic data. Thus, it was essential to develop methods to generate virtual health care records that met specific requirements and represented both 'normal' endemic visits (the background) as well as outbreak-specific records (the injects).

METHODS

To create the background data, emergency department (ED) visits were generated for each political unit in the NCR. ED data available for an entire NCR county was used to calculate age- and month-specific representative ED syndrome rates. These rates were applied to the census population for each county/city in the NCR to determine the expected number of ED visits for that region by age, month and syndrome. Computer programs were written to generate the expected number of records for each category. Age was assigned randomly to records within the specified age range. Sex and zip code were assigned randomly in the proportions observed in the NCR. All hospitals in the NCR were identified and expected cases were randomly assigned to surrounding zip codes. Hospital chief complaints (CC) categorized by age and month were culled from actual records and assigned quasi-randomly to the simulated ED records.

The exercise required the creation of virtual ED visits, military clinic office visits, and over-the-counter (OTC) drug sales for the inject data. Each simulated infected individual was represented uniquely by a software agent with properties such as age, sex, military/civilian generated via stochastic draws from probability distributions dictated by census data. A home zip code was assigned to each individual based on the population and median household income of a subset of NCR zip codes. Incubation periods were drawn stochastically from a lognormal distribution to create a believable epi-curve. Once symptomatic, each simulated individual sought medical treatment according to a heuristically derived health care utilization behavior model at a time dependent on symptom onset and duration and at a location based on the distance between the agent's home zip code and the health care facility. The results of the Matlab-based simulation were translated into health indicator record formats. Once the data was generated, internal public health experts evaluated the background data and inject data both individually and combined to ensure the utmost scenario realism.

RESULTS

A total of 2,623,226 background visit records were generated for the exercise. The male:female population ratio did not vary by age or month so sex was not factored into the representative ED syndrome rates. Initially, chief complaints were assigned without regard to gender, leading to a few records with improbable chief complaints, such as a pregnant male, that were corrected prior to the exercise. Use of the broad age categories in the NCR adversely affected the distribution of CC in some categories. For example, an unexpectedly high number of 18-30 year olds with probable cardiac illness were created.

For the injects, 4394 civilian and 106 military males became symptomatic with tularemia between 4/5-5/3/05. Those infected generated 3050 ED visit records, 106 military office visit records, and 2627 sales of OTC medication at 23 hospitals, 6 military clinics, and 453 drug stores located throughout the NCR. Infected patients were dispersed among affluent zip codes due to the expense of the targeted box seats. Apart from the mentioned minor data anomalies, the realism provided by the simulated population and outbreak data fully engaged all participants enabling a very successful exercise.

CONCLUSIONS

Outbreak simulations allow public health officials to practice and refine investigation protocols and can offer a way to evaluate and compare electronic biosurveillance systems. The methods described here can easily be applied to other populations and systems to support numerous different exercises.