Evaluating Outbreak-Detection Methods Using Simulations: Volume Under the Time-ROC Surface

Ken Kleinman, Allyson Abrams

From the Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care

OBJECTIVE

We developed metrics for evaluating tools used for outbreak detection, assuming simulated outbreaks.

BACKGROUND

There are many proposed methods of identifying outbreaks of disease in surveillance data. However, there is little agreement about appropriate ways to choose amongst them. One common basis for comparison is simulating outbreaks and adding the simulated cases to real data streams ('injected outbreaks'); competing statistical methods then attempt to detect the outbreak.

The receiver operating characteristic (ROC) curve and the area beneath it are well-known approaches to evaluation. The ROC curve plots the sensitivity against 1 less the specificity for a range of decision thresholds. Unfortunately, defining ROC curves in this context is not straightforward. In the usual setting of screening, ROC curves are constructed based on individuals, not populations, and it is unclear how to extend the concept to surveillance. In addition, the sensitivity and specificity need to be supplemented by the timeliness: a method with perfect sensitivity and specificity that detects outbreaks too late is useless.

METHODS

We define the conditional receiver operating characteristic (ROC) curve [1], as the ROC curve requiring that no outbreaks such as that generated by the simulation occur in the real data, with the following specifications. The sensitivity is defined as the probability of detecting an attack, calculated as the proportion of simulated attacks detected at a given detection threshold. One less the proportion of days with alarms in the real data is used as the specificity in the ROC curve; the condition requires that these are false alarms. Note that with injected outbreaks the sensitivity can be estimated with arbitrary precision (by simulating more releases) but the specificity can only be estimated once per day in the real data.

We developed three tools for evaluation that incorporate the timeliness of the signal. Each of the tools is a three-dimensional generalization of the ROC curve, with the third dimension representing time. The resulting surfaces we label as time-ROC surfaces and the volume under one as VUTROCS. Analogous to the area under the ROC curve, the maximum VUTROCS of 1 suggests perfect sensitivity and specificity, with all the signals generated at the first possible moment. A larger VUTROCS implies larger specificity or sensitivity or an improved timeliness, or a combination of these features. Each method can be weighted to ascribe more importance to early detection.

To avoid notation, we provide only a heuristic description of each method. 1) The 'tent' method begins with the usual conditional ROC curve. The third dimension is introduced by plotting the height of each point on the curve as the average proportion of days advance notice per signal with respect to a baseline surveillance system or constant minimum acceptable time of detection. The heights are connected across the ROC curve to form a curtain, then connected to the $(x,y,z) = (1,0,1)$ corner of the ROC curve to make a tent-like shape. 2) The 'step ROC' method calculates the ROC curve for detection by each day after the beginning of the outbreak. Each of these is assigned a width with the sum of the widths equaling 1. The VUTROCS is the trapezoidal volume under the curves when number of days after outbreak is the y-axis. 3) The 'two-threshold' method calculates the proportion of outbreaks detected by several time points and plots each of these against the sensitivity and specificity.

We assessed the performance of the VUTROCS tools in a previously published simulation setting [1].

RESULTS

Across a range of simulation parameters, the three tools were notably similar, though the step ROC method tended to generate a higher VUTROCS than either the tent or the two-threshold method. In contrast, the conditional ROC curve resulted in much bigger values as a result of its lack of a timeliness component.

CONCLUSIONS

Any of the proposed VUTROS methods will provide in a single number a sense of the key surveillance characteristics: sensitivity, specificity, and timeliness. This is a vast improvement over methods that require two values or ignore one key characteristic.

REFERENCES

[1] Kleinman KP, Abrams AM, Mandl KD, Platt R. Simulation for assessing statistical methods of bioterrorism surveillance. 2005 To appear in *Morbidity and Mortality Weekly Report*.