2008

Assessing the utility of public health surveillance using specificity, sensitivity, and lives saved

Ken P. Kleinman

Allyson Abrams

Assessing the utility of public health surveillance using specificity, sensitivity, and lives saved

Ken P. Kleinman and **Allyson M. Abrams**

Abstract

In modern surveillance of public health, data may be reported in a timely fashion, and include spatial data on cases in addition to the time of their occurrence. This has lead to many recent developments in statistical methods to detect events of public health importance. However, there has been relatively little work into methods to identify how to compare such methods. One powerful rationale for performing surveillance is earlier detection of events of public health significance; previous evaluation tools have focused on metrics which include the timeliness of detection in addition to sensitivity and specificity. However, such metrics have not accounted for the number of persons affected by the events. We re-examine the rationale for this surveillance and conclude that earlier detection is preferred because it can prevent additional morbidity and mortality. Based on this observation, we propose evaluating the number of cases prevented by each detection method, and include this information in assessing the value of different detection methods. Using this approach incorporates more information about the events and the detection and provides a sound basis for making decisions about which detection methods to employ.

Keywords

public health surveillance; spatio-temporal cluster detection; methods evaluation

---

1 Introduction

One important purpose of health surveillance is to get early warning of changes in the state of the public's health. There are methods old and new to detect anomalies in "live" data streams; these anomalies may reflect emergent public health state changes that could benefit from early intervention. For example, a population affected by anthrax would likely show excess cases of individuals with non-specific prodromal symptoms. [1,2] If these excess cases could be detected as anomalous and identified as anthrax, earlier prophylactic treatment could prevent cases of anthrax and earlier correct treatment of symptomatic cases could prevent mortality. [1]

In recent years, many groups have begun to perform health surveillance using large automated data sets. [3] Examples of this new kind of data include outpatient and emergency room visits, ambulance and 911 calls, and over-the-counter medication sales. [3] A common feature of such data sets is that their collection requires minimal human intervention: data collected electronically for another purpose are copied to surveillance systems. Another common feature is the lack of definitive diagnostic data on cases, so that systems report a count of persons buying a particular medication or sharing a common symptom.

Surveillance systems based on such data are often referred to as “syndromic surveillance systems.” The definition of syndromic surveillance could be refined, but for our purposes the above simplistic description is sufficient.

In surveillance using traditional data streams, such as passive reporting of cases or sentinel physician network reports, data have usually not been collected in automated ways, and this has limited the amount of data collected, as well as the timeliness of reporting. Specifically, data reported has often been limited to a simple case count, by week, usually for weeks that closed more than two or three days previously. These data streams have most often been assessed weekly, and have summed data from metropolitan regions, states, or whole countries. In this setting, surveillance is reminiscent of quality control, and most methods applied to such data streams are in fact appropriated directly from the quality control literature. [e.g. 4]

However, the data available in syndromic surveillance systems are frequently available daily or more often, with reporting lags no longer than two or three days and frequently measured in hours. [3] Additional data are often available with each case, the most salient of these being a zip code or census location where the case was seen, resides, or works. Many recently-proposed statistical methods incorporate this spatial data in an attempt to improve the ability to detect anomalies in the data stream. [e.g. 5–7]

However, the additional data and its use in surveillance introduce a new problem: how should system designers decide which detection method is best for surveillance? Methods from quality control can draw on the long history of evaluation methods for such tools, which describe optimal surveillance techniques when the mean and variance are constant, and when the alternative to be detected has simple properties, such as a new constant mean. [8,9] The shift to data with non-constant mean and variance (due to predictable patterns based on e.g. day of week, season, etc.) as well as to spatial settings with their non-constant expectations per region and complex alternatives render these methods inapplicable. [10]

In previous work, we proposed metrics and tools which can help compare and evaluate detection methods while accommodating various handicaps and incorporating specificity and sensitivity as well as time to detection. [10–12] In this article we introduce a new approach to such metrics which accounts for the number of people affected by the anomaly. In section 2, we describe the environment for assessment. In section 3, we describe the new metric. In section 4, we describe an example setting and show the results of the new metric. Finally, in the last section, we discuss the results.

2 Background

We assume that data exist regarding *events* that the surveillance system should detect, including the truth regarding which persons in the data stream (“cases”) were affected by the event and which are “noise.” This truth could be discovered through extensive research conducted in a real event, or known because the event has been simulated. We also assume that each statistical method used to attempt to detect the event in the data stream results, after each reporting period, in an assessment of the probability an event has occurred. For convenience, we think of this probability loosely as a p-value assessing the null hypothesis that no event has occurred, though other formulations may also be useful. [5] Denote this probability p_{mi} for method m applied to event i. Method m is said to *detect* event i at threshold t if p_{mi} is smaller than threshold t and some criteria of detection is met. An example criteria of detection is that at least one case caused by the event is in the signal. Here we assume that the detection method to be evaluated generates spatio-temporal *signals*

or regions identified as being part of the anomaly. By convention, we use days hereafter to denote data reporting periods.

For a given *t*, define the sensitivity (the probability of a positive test, given true positive) as the proportion of events that are detected; define the specificity (probability of a negative test, given true negative) as the proportion of non-event surveillance days with probability assessments greater than *t*. In other contexts plotting sensitivity vs. 1-specificity across values of *t* would generate a receiver operating characteristic (ROC) curve. In this case there is a discordance between the definitions of the two test characteristics: one is defined per event and the other per day. Together with the spatial requirement of detection, this discordance means that many results pertaining to ROC curves do not apply, and thus we refer to the curve in this setting as an ROC-like curve. [11] In particular, the observation that a useless test should have an area under the ROC curve of 0.5 does not hold here; similarly, the notion that the test could profitably be inverted if the area were less than 0.5 is not appropriate. Finally, when events are simulated and the simulated cases are superimposed on a set of observed data, the precision of the specificity as described here is limited by the length of the surveillance period (since non-events cannot be simulated) while the sensitivity can be estimated with arbitrary precision by simulating additional events.

A final feature of the setting is that many events may include cases entering the data stream during multiple days. Since data streams are assessed for anomalies once per day, a statistical detection method may detect the event on any day during which the event is ongoing. We extend the above notation as p_{mid}, where *d* indicates the day to which the method was applied; for convenience define *d* as the number of days since the beginning of the event. This feature means that we must consider not just the traditional test characteristics in the modified form described above, but also the timeliness of the detection.

All else being equal, we should prefer a method which detects an event earlier, as this will allow greater prevention of morbidity and mortality. If the event is detected, define the detect time for a given method *m* at a given threshold *t* as the minimum *d* such that $p_{mid} < t$. Seen from a different perspective, in the surveillance setting there are several tests during which an event could be detected, leading to a multiple comparisons situation in which arbitrary tests appear to detect events due to the many tests applied over the course of each event. This problem has less impact when the detection rule requires cases in the event to appear in the region of the signal; we ignore this problem for the remainder of the article.

In previous work, we considered incorporating the timeliness by weighting the plotted points on the ROC-like curve by the average proportion of time saved by the detection, relative to some observed or postulated reference signal. We also suggested two three-dimensional analogues of the ROC curve incorporating various versions of the timeliness of the detection. [10]

3 Weights by lives saved

Here, we step back from the question of timeliness and recall that the earlier signals are preferred because of the potential for earlier signals to prevent more morbidity and mortality. We begin from an assumption that all morbidity and mortality are of equal value to prevent; straightforward adjustment of the method described below can be used if it is possible to quantify the relative value of morbidity vs. mortality or of different morbidities. Starting from the principle of equal value, we develop a method to adjust the ROC-like curve by the relative value of the detection. Specifically, we consider the number of cases caused by the event that would have been prevented by a given detection. Our objective here is to give greater value to methods which detect events with more lives saved and smaller value to those detections saving fewer lives. Note that events which only affect a few cases

cannot receive a great value in this approach; this is a fundamental change in perspective, as compared to the principle of treating all events and hence all detections with equal timeliness equally.

We implement this new perspective as follows. For every statistical method *m*, event *i* and threshold *t*, we calculate the number of cases caused by the event and occurring on or after the day of detection *d*; denote this "number saved" by s_{mit}. Other definitions are also plausible: for instance, only cases which are observed within the syndromic surveillance system might be counted. We propose the weight for a given event, threshold and method be a function of s_{mit}.

One obvious function would be $f(s_{mit})=s_{mit}/k$, where *k* is a constant. An attractive value for *k* would be the maximum number of cases *ncases* caused across some group of events; this would give full weight only for a detection on the first day of cases for the largest event in that group. We refer to these weights as the "ideal" weights based on the idea that only an ideal system would achieve weights of 1 under this weight function, and only if all events affected the same number of individuals. For events with a skewed distribution of the number of cases caused by an event, the ideal weights might result in most $f(s_{mit})$ being similar. This would reduce the discriminatory ability of the eventual adjusted ROC-like curve. In such cases, we suggest a generalization such as

$$f(s_{mit})=\begin{cases} \frac{s_{mit}}{k} & s_{mit} \le k \\ 1 & s_{mit}>k \end{cases}. \qquad (1)$$

For $k = ncases$, this is identical to the ideal weights; for smaller *k* equation (1) will give a relatively larger weight to detections saving fewer cases, while still penalizing late detections or detections in smaller events. Other plausible similar functions include those incorporating a floor as well as a ceiling, or logistic functions; a different approach would use the empirical CDF of the number saved to find the weights, i.e.

$$f(s_{mit})=\sum I(s_{mit}<l)/N.$$

Whatever weight function is used results in a weight for every method at a given threshold for every event. These will be averaged across a set of events for a given threshold and method. The resulting average weight for a given threshold will be multiplied, for example, with the sensitivity achieved at that threshold to find a lives-saved-weighted sensitivity, and the ROC-like curve created we dub the "lives-saved-weighted ROC-like curve."

A notable feature of using weight functions like (1) or those alluded to above is the impact of defining the group of events which use a given *k*. If the group is too broad, there may be many weights near 0 or 1. In contrast, too many groups could make results difficult to interpret.

4 Example

As an example we apply the technique to a previously documented simulation study. [11] Briefly, we simulated anthrax dropped from the height of a cropdusting plane, with drops occurring distributed uniform in space within two regions: an urban area around Boston, MA, and a surrounding suburban region. The anthrax spores fell in one of two patterns,

referred to here as ‘Class A’ and ‘Sverdlovsk’ without further description for the sake of brevity. [2,13] The probability that exposure to a spore caused illness was set to one of 5 values: 10^{-10}, 5×10^{-10}, 10^{-9}, 5×10^{-9}, and 10^{-8}. Other features of cases, such as the time to symptoms from exposure, given disease, were based on published findings. [1] Simulated cases were added to the raw data from a surveillance data stream, detection was attempted, then the cases were removed before the next set of simulated cases were added. This process was repeated with unique simulated cases generated by anthrax dropped three times for each day of a calendar year. This resulted in $2 \times2\times5\times1095 \approx 20{,}000$ simulated events.

We applied 7 methods in attempting to detect each event. These included space-and-time scan statistics, modified to reflect variable seasonal baselines, and with a maximum signal length of 1, 3, and 7 days as well as Poisson-based generalized linear mixed model assessments with exact signal lengths of 1, 3, and 7 days. [5,14–16] The former were fit using freely-available SaTScan software. [17] We also used a purely temporal model based on a time-series regression, resembling methods proposed by Reis and colleagues. [18–19]

We defined a detection as a signal which includes cases caused by events. For the scan and mixed model approaches, this implies both that the region included in the signal as well as the time indicated by the signal included cases caused by the event. For the purely temporal method, any signal including days during the event was considered a detection; this gives some advantage to the purely temporal method.

For the weighted ROC area incorporating the lives saved, we treated each parameter combination (urbanicity, pattern, probability of illness) as a set. As weight functions, we used the ideal weights as well as weight functions of the form of (1) with k set to the median and the 75^{th} percentile of the number of cases saved.

Example results are shown in tables 1 and 2; complete results can be found in the Appendix. Table 1 shows results for the Class A spore distribution, for spores dropped in the urban region, and using the ideal weights. Table 2 shows the same set of simulated events, but using equation (1) with k set to the median number of lives saved in that set. The methods show face validity in reading across each row: the larger the probability of illness per spore, the greater the number of cases and thus the greater area. Reading down each column shows that the mixed effects models were superior to the scan or time-series approaches. In addition, the 3-day fixed signal length proved a superior detection tool among the mixed effects approaches. Comparing the two tables shows the expected result that the areas tend to be greater when k is the median number of lives saved; when k is the 75^{th} percentile, the values lie between the two examples shown.

5 Discussion

We have motivated and described a new metric for evaluating statistical signals in health surveillance. The method shown is a new approach in that it abandons timeliness of the signal as a mere proxy for the potential savings in morbidity and mortality and incorporates that savings directly. By avoiding timeliness as a separate feature it simultaneously simplifies the problem of evaluation while incorporating more information.

We demonstrated the application of the method in a simulated anthrax attack in the Boston area; the demonstration shows the method results in clear conclusions as to which detection algorithm should be preferred across the range of simulated events explored. In particular, the evaluation shows a clear preference for mixed-effects models over space-and-time scan statistics. It also shows that the time-series approach was markedly inferior to any spatio-temporal approach. Using the approach with the ideal weights lead to smaller areas, which may be a more realistic assessment of the utility of surveillance. Weight functions using

equation (1) with smaller denominators, while perhaps over-optimistic about the absolute value of the applied methods, appear to give a greater relative range. Ideal weights might be used in deciding whether to institute a surveillance system at all, while smaller denominators could be used to choose among methods to be implemented once a system was instituted.

Some other approaches to incorporating the number of persons affected bear discussion. One would be to weight by the proportion of cases detected. However, this gives a greater weight to a method which detects an event of two people early enough to save one of them than to a method which detects an event of 10,000 people in time to save 1,000. While one might prefer the former method if it were their life that were saved, without such knowledge *a priori*, the latter method would seem preferable. Another approach would be to use a similar scheme to that proposed, but to use the maximum number saved among tested methods as the denominator. While weights generated this way would have the same rank as those proposed, they would give an overly false impression of how well the methods perform in the observed setting, since they are calibrated only to the savings attainable among the methods tested, not among all methods including the ideal method.

Finally, the concept of a reference signal—a baseline method which tested methods might improve upon— could be incorporated by using the number of lives saved by the reference signal. A natural step in that case would be to compare the number saved in each method against the number saved using the reference signal, using the maximum of the quotient and 1 as the weight. This often would give higher weights to most event/method/threshold triplets if the proposed methods were better than the reference method. In that case, there would be many weights of 1, diminishing the discriminative ability of the evaluation metric. It would also give equal credit to saving more lives than the reference signal regardless of the number of lives affected. A more apt approach could be to include the reference signal with the potentially improved methods in using the lives-saved-weighted ROC-like curves.

We note that the general paradigm of using the number of lives affected directly, rather than its proxy, timeliness, can be adapted to other approaches to incorporating timeliness. For example, we have proposed using timeliness as a third dimension in generating surfaces analogous to two-dimensional ROC curves. [10] The number affected could easily be used in place of timeliness in these applications.

A common and valid complaint about ROC curves is that they give equal importance to all values of specificity. In many cases and certainly in the case of surveillance, small values of specificity are untenable: false alarms can be costly and tend to inure responders to the importance of signals. If this is a concern, the region of the weighted ROC-like curve can be truncated to just those specificities which can be accepted in the context of the application.

The proposed approach shows potential as a unified metric which can be used in evaluation. It is better than previously proposed methods in that it incorporates information on the number of people affected; previous methods examine the timeliness of the signal without assessing whether it is equally important to detect each signal in a timely fashion. Using the new metric will allow system designers to better choose which detection methods to employ.

References

1. Brookmeyer R, Blades N. Prevention of inhalational anthrax in the U. S. outbreak. Science. 2002; 295(5561):1861. [PubMed: 11884746]
2. Meselson M, Guillemin J, Hugh-Jones M, Langmuir A, Popova I, Shelokov A, et al. The Sverdlovsk anthrax outbreak of 1979. Science. 1994; 266(5188):1202–8. [PubMed: 7973702]
3. Bravata DM, McDonald KM, Smith WM, Rydzak C, Szeto H, Buckeridge DL, et al. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. Annals of Internal Medicine. 2004; 140(11):910–22. [PubMed: 15172906]
4. Hutwagner LC, Thompson WW, Seeman GM, Treadwell T. A simulation model for assessing aberration detection methods used in public health surveillance for systems with limited baselines. Statistics in Medicine. 24(4):543–550. [PubMed: 15678442]
5. Kleinman K, Lazarus R, Platt R. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. American Journal of Epidemiology. 2004; 159(3):217–224. [PubMed: 14742279]
6. Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A Space-Time Permutation Scan Statistic for Disease Outbreak Detection. PLoS Med. 2005; 2(3):e59. [PubMed: 15719066]
7. Wong W-K, Moore A, Cooper G, Wagner M. WSARE: What's Strange About Recent Events? Journal of Urban Health. 2003; 80(2 Suppl 1):i66–75. [PubMed: 12791781]
8. Frisen M, Demare J. Optimal surveillance. Biometrika. 1991; 78:271–290.
9. Frisen, M.; Sonesson, C. Optimal Surveillance. In: Lawson, AB.; Kleinman, K., editors. Spatial and Syndromic Surveillance for Public Health. Chichester; Wiley: 2005. p. 51
10. Kleinman K, Abrams A. Metrics for assessing the performance of spatial surveillance. Statistical Methods in Medical Research. 2006; 15:445–464. [PubMed: 17089948]
11. Kleinman K, Abrams A, Mandl K, Platt R. Simulation for assessing statistical methods of bioterrorism surveillance. Morbidity and Mortality Weekly Report. 2005; 54(supp):101–108. [PubMed: 16177700]
12. Kleinman K, Abrams A, Yih WK, Platt R, Kulldorff M. Evaluating spatial surveillance: detection of known outbreaks in real data. Statistics in Medicine. 2006; 25:755–769. [PubMed: 16453375]
13. Spijkerboer HP, Beniers JE, Jaspers D, Schouten HJ, Goudriaan J, Rabbinge R, et al. Ability of the Gaussian plume model to predict and describe spore dispersal over a potato crop. Ecological Modeling. 2002; 155:1–18.
14. Kulldorff M. Prospective time periodic geographic disease surveillance using a scan statistic. Journal of the Royal Statistical Society, Series A. 2001; 164:61–72.
15. Kleinman K, Abrams A, Kulldorff M, Platt R. A model-adjusted space-time scan statistic with an application to syndromic surveillance. Epidemiology and Infection. 2005; 133:409–419. [PubMed: 15962547]
16. Kleinman, K. Generalized Linear Models and Generalized Linear Mixed Models for Small-Area Surveillance. In: Lawson, AB.; Kleinman, K., editors. Spatial and Syndromic Surveillance for Public Health. Chichester; Wiley: 2005. p. 77-94.
17. Kulldorff, M. SaTScan: Software for the spatial and space-time scan statistics. [Accessed September 28, 2007]. www.satscan.org
18. Reis BY, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. Proc Natl Acad Sci U S A. 2003; 100(4):1961–5. [PubMed: 12574522]
19. Reis BY, Mandl KD. Time series modeling for syndromic surveillance. BMC Med Inform Decis Mak. 2003; 3(1):2. [PubMed: 12542838]

Table 1

Area under the lives-saved-weighted ROC curve with spores falling in the Class A shape in an urban location and with $k = ncases$ in equation (1) for each of the 7 methods

Method	**Pr(Illness) =10^{-10}**	**Pr(Illness) =5×10^{-10}**	**Pr(Illness) =10^{-9}**	**Pr(Illness) =5×10^{-9}**	**Pr(Illness) =10^{-8}**
GLMM1[a]	0.10231	0.10211	0.12168	0.28881	0.36812
GLMM3[a]	0.05257	0.05800	0.08251	0.30608	0.37462
GLMM7[a]	0.02765	0.03695	0.05904	0.27755	0.34064
Scan1[b]	0.00005	0.00055	0.00289	0.04148	0.06062
Scan3[b]	0.00020	0.00312	0.01090	0.10713	0.13877
Scan7[b]	0.00051	0.00665	0.01849	0.16065	0.20426
Time-series	0.00049	0.00304	0.00462	0.01054	0.01611

[a] Mixed effects Poisson models with signal lengths of exactly 1, 3, and 7 days

[b] Space-and-time scan approaches with maximum signal lengths of 1, 3, and 7 days

Table 2

Area under the lives-saved-weighted ROC curve with spores falling in the Class A shape in an urban location and with the weight function as in equation (1) with k = the median cases saved for each of the 7 methods

Method	Pr(Illness) = 10^{-10}	Pr(Illness) = 5×10^{-10}	Pr(Illness) = 10^{-9}	Pr(Illness) = 5×10^{-9}	Pr(Illness) = 10^{-8}
GLMM1[a]	0.25034	0.25620	0.29364	0.67661	0.83211
GLMM3[a]	0.12863	0.14552	0.19910	0.71707	0.84680
GLMM7[a]	0.06766	0.09270	0.14247	0.65024	0.77000
Scan1[b]	0.00012	0.00139	0.00698	0.09718	0.13704
Scan3[b]	0.00049	0.00784	0.02629	0.25098	0.31368
Scan7[b]	0.00125	0.01669	0.04463	0.37638	0.46172
Time-series	0.00120	0.00762	0.01115	0.02469	0.03641

[a] Mixed effects Poisson models with signal lengths of exactly 1, 3, and 7 days

[b] Space-and-time scan approaches with maximum signal lengths of 1, 3, and 7 days