2006

Evaluating spatial surveillance: detection of known outbreaks in real data

Ken Kleinman

Allyson Abrams

W. Katherine Yih

Richard Platt

Martin Kulldorff

Evaluating spatial surveillance: detection of known outbreaks in real data

Ken Kleinman, Allyson Abrams, W. Katherine Yih, Richard Platt
and Martin Kulldorff

SUMMARY

Since the anthrax attacks of October 2001 and the SARS outbreaks of recent years, there has been an increasing interest in developing surveillance systems to aid in the early detection of such illness. Systems have been established which do this is by monitoring primary health-care visits, pharmacy sales, absenteeism records, and other non-traditional sources of data. While many resources have been invested in establishing such systems, relatively little effort has as yet been expended in evaluating their performance.

One way to evaluate a given surveillance system is to compare the signals it generates with known outbreaks identified in other systems. In public health practice, for example, public health departments investigate reports of illness and sometimes track hospital admissions. Comparison of new systems with extant systems cannot generate estimates of test characteristics such as sensitivity and specificity, since the actual number of positives and negatives cannot be known. However, the comparison can reveal whether a new or proposed system's signals match outbreaks detected by the existing system. This could help support or reject the new system as an alternative or complement to the extant system.

We propose three methods to test the null hypothesis that the new system does not signal true outbreaks more often than would be expected by chance. The methods differ in the restrictiveness of the assumptions required. Each test may detect weaknesses in the new system, depending on the distribution of outbreaks and can be used to construct confidence limits on the agreement between the new system's signals and the outbreaks, given the distribution of the signals. They can be used to assess whether the new system works in that it detects the outbreaks better than chance would suggest and can also determine if the new systems' signals are generated earlier than an extant system.

KEY WORDS: surveillance; spatial methods; evaluation; permutation tests; scan statistic

INTRODUCTION

Since the bioterrorist attacks using anthrax in October 2001 and the SARS outbreaks in subsequent years, there has been increased interest in surveillance to detect such outbreaks. Traditional public health surveillance generally relies on post-diagnosis reporting of cases by care providers. Many resources have been devoted to systems that attempt earlier detection by collecting data on the patterns of illness in the community [1–7]. For naturally occurring disease, these general patterns may provide an earlier warning than other systems; for bioterrorist attack, an unusual number of sick individuals may be the first sign of an attack, barring detection of the bioterrorist agent in the environment.

The systems have many aspects, ranging from the way data is collected to how it is analysed and the responses generated if the analysis proves alarming [8]. While some guidance on the construction of such systems is available, there is little specific information on how to assess them. Such evaluation can be based on the retrospective 'performance' of the system in the presence of real outbreaks, on fully simulated data, or on simulated attacks laid over real historical data. In the present article, we consider the case of how the system would have performed at detecting outbreaks known to have occurred in the past. Due to the happy dearth of known bioterrorist attacks, we provide an example based on known outbreaks of gastrointestinal illness. However, the statistical approach we take is general with respect to the kind of outbreaks that are used for comparison.

On the other hand, we limit our interest to systems that contain spatial (geographical) data on the cases [4, 6]. These systems, when coupled with statistical techniques that use the spatial data [9–11], hold out the promise of greater sensitivity and specificity than is possible in methods that ignore geographic location [10].

There are two substantive questions we wish to address in this context: (1) Can the surveillance system detect known outbreaks? (2) Can the surveillance system detect known outbreaks before they would otherwise have been detected? Both questions speak to the value of a proposed system relative to the method that identified the real outbreaks. If the new system can detect the outbreaks well, it might be considered as a replacement system, and would certainly be viewed as having some face validity with respect to these real results. Similarly, earlier detection of real outbreaks would improve public health to some degree and would point to the value of a new system. The proposed methods address both questions.

In the methods section, we first lay out some notation, including formalizing the notion of a 'hit' as a signal from a surveillance system that can be related to a real outbreak. We then motivate and describe three tests of the null hypothesis that the observed number of hits is no more than chance association between signals and outbreaks, given the distributions of outbreaks and signals. The goal of assessing the co-location of two series of points in space and time has been addressed previously by Klauber [12]. One of our proposed tests resembles a special case of the framework proposed there; the other two tests lie outside it. In addition, our setting introduces events that contain multiple points in space and time. Finally, the application to signals generated from surveillance is novel.

In the results section, we first provide a simple artificial data example in which various violations of the assumptions needed for the tests may be violated, demonstrating that the tests can be misleading if applied without examining the assumptions carefully. We then show an example in real data. In the discussion section, we examine the utility of the tests and make observations on their limitations. We also outline the method by which definitions of a 'hit' can be altered to address the question of the timing of signals.

METHODS

Suppose that surveillance data is available over some time period comprising days $1,\ldots,T$. Also, suppose that in the same period N_{real} 'real' outbreaks have been confirmed by an epidemiologically accepted method. Each of the outbreaks also has a location, typically the site of the exposure to a disease-causing agent. This could be an arbitrary region, a zip code, a census tract, or a latitude and longitude pair. Arbitrarily index the outbreaks $i = 1,\ldots,N_{\mathrm{real}}$, and let $\mathrm{REALLOC}_i$ and $\mathrm{REALSTART}_i$ designate the location where and time when outbreak i began. Note that outbreaks often continue for more than one day; let $\mathrm{REALEND}_i$ be the last day of outbreak i.

Now, suppose that a new system has collected data over the same period; assume the system includes some specific outbreak-identifying statistical analysis. To clarify the distinction between the real outbreaks and the putative outbreaks identified by statistical analysis of the new data, we refer to the latter as signals generated by the system. To best emulate the performance that would have been seen had the system been used contemporaneously as the data was collected, a statistical technique must be applied for each day t, using only data from days $1,\ldots,t$ to signal outbreaks, as opposed to including 'future' days $t+1,\ldots,T$ when looking for outbreaks on day t.

Each signal found in the new data would have a location as well. Label the signals $s = 1,\ldots,N_{\mathrm{stat}}$ with locations $\mathrm{SIGNALLOC}_s$ and start date $\mathrm{SIGNALSTART}_s$. Note that with some statistical techniques, the location may be a single zip code or census tract while others generate some geometric or general shape with a focus at some map point. As with the outbreaks, signals may identify a suspected anomaly having occurred across a period of time. Of course, the signal itself is generated at a single time point, but the putative outbreak so identified may have a depth in time; denote the end of the signal as $\mathrm{SIGNALEND}_s$.

With this notation established, we can proceed to evaluate the performance of the new system. Define a 'hit' for each signal as the event that the signal overlaps both spatially and temporally with at least one real outbreak, though, of course, other definitions are defensible. Symbolically, a hit for signal s is defined as

$$\exists \mathrm{outbreak}_i \left[\begin{array}{l} ([\mathrm{REALSTART}_i, \mathrm{REALEND}_i] \cap [\mathrm{SIGNALSTART}_s, \mathrm{SIGNALEND}_s]) \neq \emptyset \\ \text{and } (\mathrm{REALLOC}_i \cap \mathrm{SIGNALLOC}_s) \neq \emptyset \end{array} \right]$$

using the standard set and logic notation for 'there exists' ($\exists$), 'intersect' ($\cap$) and the empty set ($\emptyset$). Let $\mathrm{HIT}_s = 1$ if the signal has a hit and 0 if not. Then the performance of the system with respect to its identifying outbreaks also found by the external data could be measured as $\mathrm{TOTALHITS} = \sum_s \mathrm{HIT}_s$.

However, this statistic is not of much use in a vacuum. To demonstrate this, consider a proposed system that generated a large number of signals, each with duration over days 1 to T and with location including the whole study region. Then the number of hits would necessarily equal the number of signals without suggesting actual practical utility. Relatedly, one should consider the specificity of the signals, equivalent to the proportion of false positive signals generated. To determine the value of the statistical method, we need some sense of the variability and expectation of this statistic. However, there is no obvious way to obtain these.

We provide three alternative randomization-based tests that can demonstrate whether the observed number of hits is more or less than would be expected if the signals had no relationship with the outbreaks. Note that we do not address here the important question of how to compare statistical signal-generation methods. This question is complicated by the possibility of different numbers of signals generated by different methods and by the incompleteness of the known outbreaks with respect to all outbreaks.

Test I: *random signals*

Suppose the statistical method employed by the new system was limited to generating signals of identical duration and size. In addition, suppose that the population was distributed uniformly across space and more specifically that there was no spatial pattern to the outbreaks. Suppose that there was no seasonal pattern to the outbreaks, and that they were also of a single duration.

In that case, we could generate the null distribution for the number of hits by choosing N_{stat} points from a two-dimensional uniform distribution and choosing for each point a random (discrete uniform) date from the surveillance period. We would then declare a hit for these randomly generated signals as we did for the signals generated by the surveillance system. Summing the hits for the random signals would give one instance of the number of hits observed under the null hypothesis. We would repeat this experiment many times, and thus discover the probability of 0 hit, 1 hit, 2 hits,..., N_{stat} hits under the null hypothesis that signals were randomly assigned to the space–time universe of the surveillance period. Using this, we could express a *p*-value for the observed number of hits in the real data as the probability that as many or more hits would be found under the null hypothesis.

Unfortunately, these conditions will rarely, if ever, be plausible, and the test will have little value if applied when they are violated. For example, suppose the population were unevenly distributed on the map, as would happen if the region included urban and suburban or rural areas. In that case, signals that included the urban region might be more likely to overlap with outbreaks. The random signals would include the urban region only in the proportion of the surveillance region covered by the urban area. Thus, a statistical test that included the urban region preferentially without reference to the observed data—for example, by preferentially selecting areas with high population density—would falsely appear to have more hits than expected under the null. Violations of the other assumptions would have similar anticonservative effects on the supposed null distributions.

Test II: *random dates for observed signals*

Seen another way, the problem with violating the assumption of uniform population distribution is that too many of the random signals land in uninteresting places; systems under consideration may inadvertently or intentionally take advantage of this by specifying signals to appear in more interesting places with greater frequency.

Suppose instead that, more realistically, the outbreaks were not distributed uniformly in the surveillance area, but maintain the assumption that there is no seasonality to the outbreaks. We could then take every signal location and assign to it a random (discrete uniform) date on the calendar. Having assigned random dates to the signals, we could then count the number of hits, as above, generating one instance of the number of hits found under the null. Then, as above, we would repeat this many times, and discover the null distribution of the number of hits that happen to occur if the observed locations were chosen.

In this case, too, an anticonservative bias may be introduced if the assumption of no seasonality is violated. To see this, imagine that there is seasonality in the data, as would be the case for lower respiratory complaints, for example [3]. Then if a statistical test generated more signals in the seasons when the condition of interest was more prevalent, the null could appear to be rejected without any real ability of the method to find an outbreak on the map. For example, a cu-sum method [2] that did not account for seasonality would tend to generate signals in seasons with larger numbers of cases. If one wanted to test that method in a context with spatial data, one might define the spatial extent of such cu-sum signals to cover the whole surveillance region. In this case the test described would spuriously suggest the system has more hits than expected under the null distribution.

Test III: *permutation test*

Permutation tests generally refer to exact tests that calculate the test statistic under every possible arrangement of the data. These arrangements would in this case be used to generate the distribution of the number-of-hits test statistic under the null hypothesis. The proportion of arrangements with as many or more hits than in the observed data would be the *p*-value, as above. A closely related approach uses many Monte Carlo re-arrangements of the data rather than enumeration of all of the possible arrangements [13]. If the number of Monte Carlo re-arrangements is 'large', and each arrangement has equal probability in each Monte Carlo replicate, then the exact and Monte Carlo permutation tests will have very similar results; the Monte Carlo permutation test is asymptotically equivalent to the exact permutation test [14]. (We do not discuss what 'large' means in the current case.) Monte Carlo approaches are useful if it is difficult or impractical to enumerate all possible arrangements of the data.

The difficult part of this process is defining what constitutes an arrangement of the data. Good [15] describes the canonical experimental case as 'losing the labels' that identify which samples are experimental and which are control cases. We define an arrangement as the pairings of outbreak locations with outbreak dates. Every arrangement would consist of pairing each date of an outbreak with the location of some possibly different outbreak. The different arrangements would include all possible pairs of dates with locations. The permutation version of this test was proposed in a generic form by Klauber [12], which considers only point-located events but is otherwise similar to the current proposal.

Note that under this test, neither a system that preferentially signals in more densely populated areas nor one that preferentially signals in outbreak-heavy seasons will falsely appear to reject the null. For signals that preferentially favour urban regions or seasons with disproportionate numbers of outbreaks, both the observed data and the permutations will result in hits at those times or places. This means the null distribution will have a relatively large probability of having as many hits as were observed. Similarly, if a statistical method were constructed that preferred urban regions in particular seasons, the permutation test would find the observed number of hits to be typical, not unusual.

There are $N_{stat}!$ possible permutations (approximately 10^{64} if 50 signals), so enumeration in this case will often be implausible, and Monte Carlo testing is recommended. The Monte Carlo approach would begin by making a list of the dates of outbreaks and an unlinked list of locations of outbreaks. It would proceed in each Monte Carlo iteration to re-order the dates of the outbreaks in a random fashion and attach them to the list of observed locations (which could be kept in their original order). Then the number of signals with hits in the outbreaks with randomly paired dates and locations would be recorded. As above, this would be repeated many times. After many iterations, the null distribution of the number of hits would be obtained.

In the above discussion, we refer to dates and locations, ignoring the duration and size aspects of outbreaks. In theory, these could also be permuted, meaning that each arrangement would include a date matched to a random duration, a random location, and a random size. We do not recommend this; if a system always generated larger sizes for urban locations or longer durations during seasons with more outbreaks, the null distribution generated using this permutation would wrongly appear to result in fewer hits than the statistical method oftener than appropriate. Instead, we include duration with date and size with location. These pairings make sense, as they keep the temporal features of outbreaks linked and the spatial aspects linked as well.

A more formal discussion of the test may be warranted. Our test statistic is the number of hits among a given set of signals. The null hypothesis is that the system performs no better than random chance, defined as applying the signals to outbreaks with the same marginal temporal and spatial distributions of the outbreaks but with no conceivable relationship to the signals. Under the null hypothesis, the arrangement we happened to observe in the real data is typical of arrangements with those temporal and spatial features; if that is the case, then the proportion of arrangements resulting in as many or more hits will be large. Under the alternative, the system performs better by using the data. In that case, few of the arrangements will result in as many hits.

RESULTS

Here, we first provide a conceptual example based on applying each test to artificial data, where exact calculations of p-values are possible, under various violations of the required conditions. We then show an example of applying tests to real data collected in Minnesota.

Example 1: artificial data

Imagine a region made up of 4 squares of equal size. Let us consider outbreaks and signals of one day and one square only. Now, suppose a surveillance period of 10 days. We will consider 5 outbreaks distributed across space and time in three patterns. In all cases, there is one outbreak on day 8 in square 3. The other 4 outbreaks are as follows: Pattern 1: they are concentrated so that all 4 occur in the first square, on days 1, 4, 7, and 9, as might occur if the vast majority of the population lived in square 1, and thus violating the assumption needed for test I. Pattern 2: they are concentrated so that all 4 occur on the first day, one in each square, violating the lack of 'seasonality' assumption needed for tests I and II. Pattern 3: they are distributed so that one occurs in square 1 on day 1, one on day 2 in the second square, 1 on day 3 in the third square, and one on day 10 in the fourth square, not violating any assumptions of any of the tests.

Table I. The three outbreak patterns, methods, and tests.

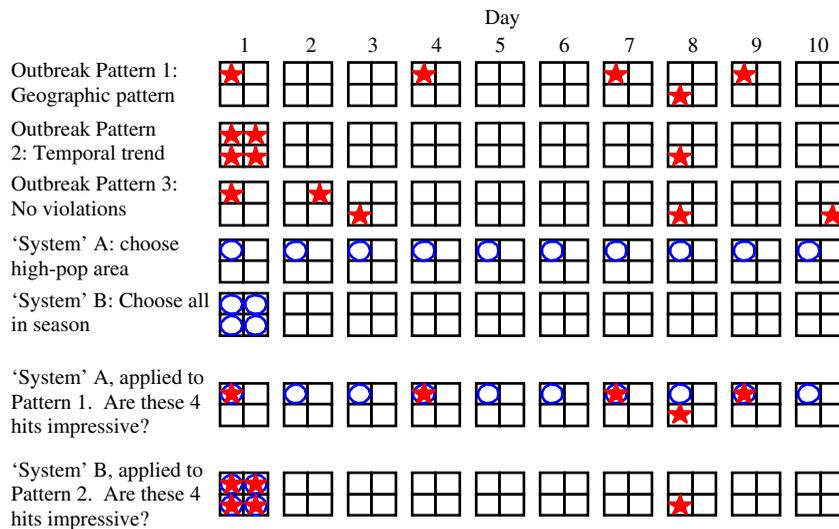|	Outbreak patterns		Systems		Tests
1	First 4 days, square 1, plus day 8 square 3	A	Square 1, every day	I	Random allocation of outbreaks to square-days
2	All four squares, day 1, plus day 8, square 3	B	Every square, day 1	II	Random dates for each outbreak
3	Three different days for square 1, 2, and 4, two other days for square 3	C	Always finds every outbreak	III	Permute location/size of outbreaks with date/duration of outbreaks

Figure 1. Schematic diagram of the artificial data. Stars are outbreaks; circles are signals. Stars appearing within circles are hits for the statistical system. Surveillance system 3 is not shown: regardless of outbreak pattern, it includes only circles around all stars. Square 1 is in the upper left-hand quadrant and the remaining squares are numbered clockwise.

Now, consider the following 'surveillance systems': System A: the system generates a signal every day in the first square and no other. System B: the system generates a signal in each square on the first day and no other. System C: the system always signals all outbreaks. Note that systems A and B have no relationship to the data, though they will sometimes get lucky and will also do quite well if they preferentially signal in areas or times when and/or where outbreaks are likely to appear. For example, an unscrupulous surveillance system designer could design a system with a signal every day in the most populous neighbourhood of those under surveillance; this would be like system A. Similarly, a system could be designed that generated a respiratory signal in every neighbourhood during the flu season; this would be like system B.

We summarize the outbreak patterns, systems, and the three tests in Table I and Figure 1. In this simplified example, we can easily derive the asymptotic or theoretical results of each

Table II. Three 'systems' evaluated by the three proposed tests. *Bold* font indicates a system for which at least one test's assumptions *are* violated by the outbreak pattern, as well as the tests the assumptions of which are violated. *Bold Italic* font indicates tests which *do not* require the violated assumptions. System A signals in square 1 each day. System B signals on day 1 in each square. System C correctly signals all outbreaks. Test I requires uniform outbreak distribution in space. Test II requires uniform outbreak distribution in time. Test III has no assumptions. All outbreak patterns contain 5 outbreaks; details are included in the table.

	Observed hits	Test I	Test II	Test III
Outbreak pattern 1: 4 outbreaks in square 1 plus one in square 3 on day 8				
System A	4	$P = \mathbf{0.027}$	$P = \mathbf{\mathit{0.618}}$	$P = \mathit{1}$
System B	1	$P = 0.414$	$P = 0.460$	$P = 1$
System C	5	$P < 0.0001$	$P = 0.0026$	$P = 0.2$
Outbreak pattern 2: outbreaks in each square on day 1 plus one in square 3 on day 8				
System A	1	$P = 0.737$	$P = 0.651$	$P = 1$
System B	4	$P = \mathbf{0.0002}$	$P = \mathbf{0.0002}$	$P = \mathit{1}$
System C	5	$P < 0.0001$	$P < 0.0001$	$P = 0.2$
Outbreak pattern 3: outbreak in square 1 on day 1, no other outbreaks in square 1 or on day 1				
System A	1	$P = 0.737$	$P = 0.651$	$P = 1$
System B	1	$P = 0.414$	$P = 0.4168$	$P = 1$
System C	5	$P < 0.0001$	$P < 0.0001$	$P = 0.017$

test. Details are presented in the Appendix. In Table II we report the number of observed hits and the probability of this number of hits under each of the three tests for each 'system'. The table demonstrates several interesting features. First of all, note that the methods that are not based on the data perform well when the outbreaks violate the assumptions of the test. System A has a small p-value (0.027) for test I, when the outbreaks have a geographic pattern as in outbreak pattern 1. System B has a small p-value (0.0002) under test II when the outbreak is focused in time, as in outbreak pattern 2. These values are bolded in the tables.

The random date test (test II) successfully protects against the spatial pattern of outbreak observed in outbreak pattern 1, returning a p-value of 0.618. In contrast, the permutation test (test III) protects against both kinds of violations of the null. Test III suggests little evidence against the null has been found, meaning that the observed number of hits in these systems and any of the outbreaks is extremely likely. These values for the less restrictive tests are in bold and italic in the tables.

Another feature demonstrated by the example is that test III may lack power in this context, even when the statistical method performs perfectly. In fact, if the 5 outbreaks all happen to occur in the same region, test III will return a p-value of 1. The minimum p-value for test III in this setting is 0.017, observed in outbreaks patterns such as pattern 3, when all four regions have signals on different days, and the fifth signal (which must appear in a duplicate region) occurs on a fifth day. In addition, for system A applied to outbreak pattern 1, test II returns a smaller p-value than test III. Since all of the assumptions for test II are met, this is a sign that test II, with more assumptions, has greater power than test III, which is more (unnecessarily) robust.

Example 2: *gastrointestinal illness surveillance in Minnesota*

As part of a nationwide system, the National Bioterrorism Syndromic Surveillance Demonstration Project, data is collected in Minnesota regarding the number of gastrointestinal complaints recorded at visits to primary care providers among a defined set of individuals [16]. Reports are summarized daily by zip code. We used data collected after 1/1/2001 to attempt to find outbreaks between 1/2/2001 and 31/1/2003. Our statistical technique used to generate signals in this system was a space-and-time scan statistic [11] adjusted to remove the effects of season, weekday, and other temporal trends [17]. We implemented the scan statistic using the freely available SaTScan^{TM} software (www.satscan.org). SaTScan returns a centroid, radius (R), and p-value for the most unusual cluster of cases. As signals, we consider only case clusters with p-values below 0.04, i.e. clusters that should be expected by chance only once in every 25 days [10] though this may not be practical in some surveillance settings. It is possible to generate signals of multiple days' length, i.e. with duration backwards in time from the day of signalling; here we allowed only single-day signals. There were 33 such signals, about twice as many as expected under the null.

The Minnesota Department of Health (MDH) is responsible for identifying and investigating outbreaks of illness in the state, including gastrointestinal illness caused by food- and water-borne organisms. The MDH provided us with a list of outbreaks investigated between 1/2/2001 and 31/1/2003, as well as the date the MDH was notified that there might be an outbreak to investigate, the date of the retrospectively identified first case, and the location (zip code) of the outbreak [18]. We omitted those outbreaks including fewer than 5 cases, those occurring in institutional environments, and those occurring outside the catchment area of the data collection. The data would be extremely unlikely to detect these outbreaks, and any observed agreement between the outbreaks and signals in those cases would almost certainly be spurious. There were 71 remaining outbreaks. Note that these are unlikely to include all of the outbreaks meeting our criteria.

We define a hit for a given signal as the case that an outbreak centroid is within $R + 25$ kilometers of the SaTScan centroid on a date between 7 days before the first known exposure and 7 days after the MDH began its investigation. These choices were made *a priori*, for the purposes of example only.

There were 22 hits. The null distribution of the number of hits under tests II and III is shown in Table III and one year of surveillance is summarized in Figure 2. Test I cannot be applied since the signals are of varying radius and the outbreaks of varying duration. The table shows that the p-value for the observed number of hits is 0.11 under test II, and 0.42 under test III. Since the assumptions required for test II to be valid are violated, it is unsurprising that the p-value observed for test II is much smaller than under test III. To be explicit, there are spatial patterns to the population density and seasons when more outbreaks occur. Thus, the test that selects random dates for the outbreaks results in a purported null distribution that contains fewer hits than appropriate. Under the appropriate test, test III, there is little evidence to reject the null hypothesis that the observed hits merely reflect a plausible number to be obtained with outbreaks with the same marginal temporal and spatial features of the observed outbreaks.

We emphasize here that it is not only the statistical method but the whole surveillance system, including the population covered, coding method, syndrome construction and innumerable other features, that is being assessed here, and an appropriate inference based on

Table III. Random date and permutation approach to generating null distributions with the proportion of randomized data sets with that number of hits the and associated *p*-value for the observed range of hits across the simulations. The italic row highlights the observed number of hits from the example.

Number of hits	Test II		Test III	
	Proportion randomizations	*p*-value	Proportion randomizations	*p*-value
10	0.0046	1.0000	0	1
11	0.0035	0.9954	0	1
12	0.0081	0.9919	0	1
13	0.0277	0.9838	0	1
14	0.0358	0.9561	0	1
15	0.0589	0.9203	0	1
16	0.1028	0.8614	0.0021	1
17	0.1351	0.7587	0.0043	0.9979
18	0.1132	0.6236	0.0461	0.9936
19	0.1628	0.5104	0.1104	0.9475
20	0.1455	0.3476	0.1758	0.8371
21	0.0901	0.2021	0.2390	0.6613
22	*0.0485*	*0.1110*	*0.1994*	*0.4223*
23	0.0381	0.0635	0.1404	0.2229
24	0.0162	0.0254	0.0686	0.0825
25	0.0058	0.0091	0.0129	0.0139
26	0.0012	0.0035	0	0.0011
27	0.0023	0.0023	0.0011	0.0011

this test might be that the surveillance system, as implemented with the described space–time scan statistic may be no better than chance. Different inferences might result when using other statistical methods or even the space–time scan statistic with different parameters, as well as more obviously with different surveillance data sources, in different time periods, in different geographical regions, applied to different syndromes, or compared to other sets of outbreaks.

DISCUSSION

We have described three tests that can be used to assess whether surveillance systems—including both data collection and data analysis features—perform better than random chance in generating signals from real data. The tests have different assumptions regarding the underlying spatial and temporal distribution of outbreaks; each can be appropriately used when their assumptions are met. We provided a detailed discussion of when such tests can be appropriately used and the implications for power of the choice of test. We see these tests as useful when evaluating syndromic surveillance systems.

   It may be worthwhile to note the distinctions between the proposed permutation test and a recently proposed space–time permutation version of the scan statistic [19]. While the proposed permutation test involves a permutation over space and time, is applied to surveillance data, and is used in the example to evaluate a system which uses a space and time scan statistic, the current application is wholly distinct from the space–time permutation scan statistic. The
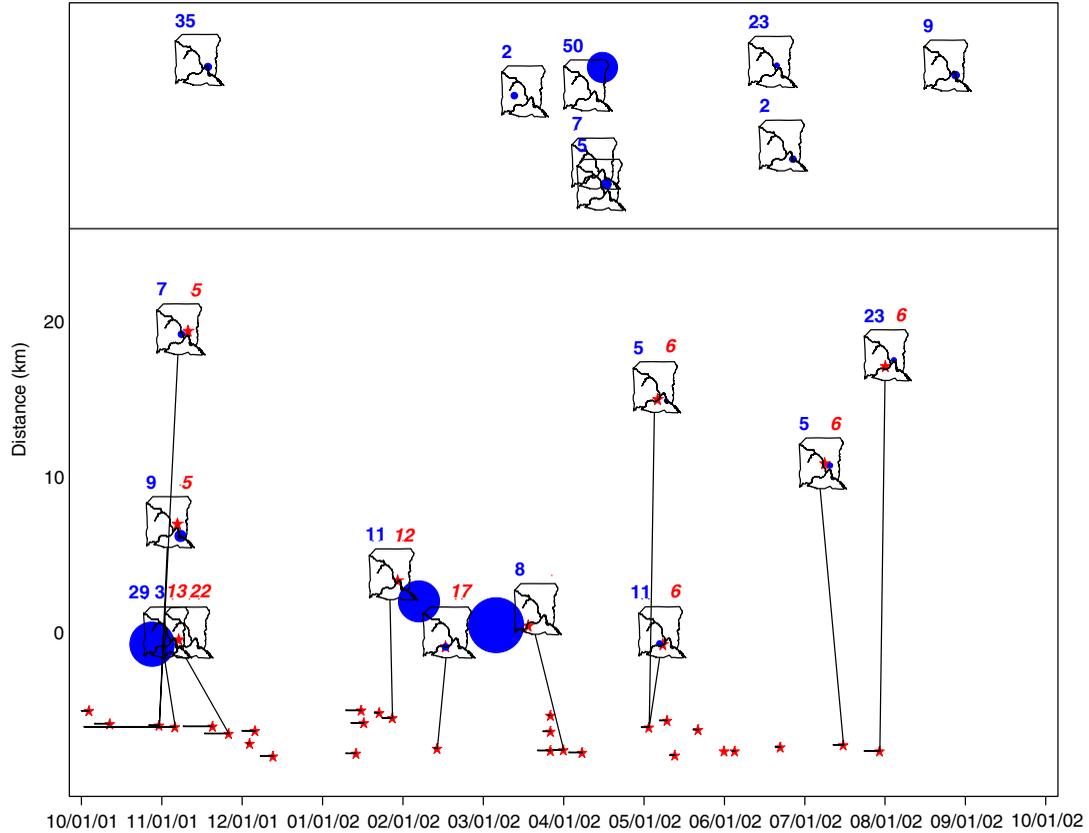
Figure 2. Relationship between known gastrointestinal outbreaks and gastrointestinal signals in health plan data, Minnesota, 2001–2002. The horizontal line running the width of the plot separates signals that are hits (below) from signals that are not (above the line, jittered vertically). Hit criteria are described in the text. Stars across the bottom of the plot represent investigation dates for outbreaks described in the text. Short horizontal lines extend from the retrospectively determined date of first exposure to the date of investigation. The small maps represent the core of the catchment area, around Minneapolis-St. Paul. The maps are placed horizontally at the date the signal was generated and vertically at the distance from the outbreak to the signal which hit it. Within each map, the dot represents the area of the SaTScan signal. If the SaTScan area had a radius of zero (i.e. consisted of a single zip code), a small circle is placed. The star in the map indicates the centroid of the zip code associated with the outbreak. Occasionally, symbols appear outside the core area; they represent events in the larger catchment area and are located accurately with respect to the corresponding map. Numbers of cases are shown above each map, black for the cases in the signal and grey for the outbreak. A hit is indicated by a line. If a linkage line's slope is negative, the signal appeared before the date the outbreak investigation began. The most negative slopes are for linkages in which the signal was both early and close in distance.

space–time permutation scan statistic [19] describes a method for performing scan statistics to identify clusters in the absence of an observed denominator, and has nothing to do with assessing the co-location of two sets of events in space and time. In fact, the space–time permutation scan statistic (or any other cluster identification tool) could be used as the signal-generation method that would later be evaluated with the current permutation test.

The conceptual example shows the effects of violating the various assumptions can be anti-conservative in that they can tend to generate p-values smaller than appropriate. In addition, it shows that the power of the most general test may be small in comparison with a test that uses the assumptions, when the assumptions are tenable. The gastrointestinal data demonstrates an actual application in public health surveillance and reinforces the negative effects of violating the assumptions.

The nature of a discrete statistic like the count of hits is that the difference between a statistically significant number of hits and an ordinary number of hits is one hit. As a conceptual example only, imagine that a system generated 50 hits, with a p-value of 0.0015. This would allow one to reject the null and conclude enthusiastically that the system's signals hit outbreaks better than random chance. But the test might also show that 49 hits would have had a p-value of 0.4. If that were the case, the system might seem less attractive.

We would like to be able to place confidence limits around the number of hits that should be expected if the method works. In contrast, the tests described above can provide only confidence limits for the number of hits that should occur under a given distribution of signals, under the null.

Finally, some comment on the use of test III may be warranted. With a small number of signals, Monte Carlo methods may not be required. A rough rule is that with 6 or fewer signals, exact calculations would be preferable, since there are only 720 permutations. Using exact calculations in these cases would also protect one from inadvertently ascribing more precision to the p-value than can actually exist.

Note that some latitude in defining a 'hit' may be appropriate. For example, it is possible that the statistical method could detect an outbreak earlier than traditional surveillance. This would be the case if the traditional surveillance only provides the date of the first *ascertained* case as opposed to the timing of the initial *possibility of exposure* to the agent. In this case, one could define the duration of the outbreaks to extend further back into the past than the first ascertained case.

One could define a useful statistical method as one which can detect outbreaks before traditional methods. In that case, one would define the duration of outbreaks as lasting only from the initial possibility of outbreak to the time that the public health department was notified of the outbreak. This would reflect the reality that syndromic surveillance systems will never replace traditional surveillance in public health practice, the purpose of the analysis being to demonstrate that statistical analysis can aid traditional surveillance by occasionally providing earlier warning of outbreaks for which public health departments may later identify causative organisms.

However, these tests do not address three important features of traditional surveillance and new systems. First, there is the question of what proportion of the outbreaks have been detected by the system. A system that identifies 10 per cent of outbreaks and a system that identifies 100 per cent of outbreaks will have different practical utility, and either or neither may provide signals that allow a rejection of the null. Nor does the tool address the question of the proportion of false positives—signals which are not hits. These may be so numerous as to preclude the use of the tool even if it is shown to be useful via a test suggested above. Finally, signals that are 'misses' may reflect true events that went undetected by whatever mechanism generated the outbreak list.

These are issues of cost and benefit that are essentially non-statistical in nature. They serve as a reminder that in public health practice, statistical results are just one source of

information in the decision-making process. In addition, however valuable a system may be, merely attaching a signal to an outbreak is not equivalent to determining the source of an outbreak, or even finding cases caused by that outbreak. The ultimate utility of a system is shown only if signals it generates can be linked to exposures to disease organisms of public health significance.

In conclusion, the presented tests can help public health officials determine whether a particular surveillance system is better than randomly suggesting signals without reference to the data. By redefining the end of the outbreak as described, the test can be used to address the question of whether the system can detect outbreaks before they would otherwise have been detected. The tests describe the behaviour of the system if random outbreaks can be assigned uniformly over space and time (test I); uniformly over time, conditioning on different probabilities of outbreak over space (test II); or with the same marginal spatial and temporal distributions as seen in the data (test III). The test results alone cannot recommend a particular surveillance system as having any great utility in practice, but they can rule out systems that can neither replicate nor improve on traditional surveillance.

APPENDIX A: PROBABILITIES OF NUMBERS OF HITS IN THE ARTIFICIAL DATA EXAMPLE

For the purposes of calculation, it is more economical to discuss the probability calculations by test than by outbreak pattern. Recall that under 'statistical' system A, there are 10 signals, under B, 4 signals, and under C, 5 signals, regardless of the outbreak pattern. In addition, it may be useful to consider randomization or permutation of the signals rather than the outbreaks. These are equivalent.

Test I

Under test I, the number of hits is distributed binomial (p, N), with parameters $p = 0.125$ ($= 5/40$, the proportion of square-days with outbreaks) and $N =$ number of signals. Thus, for system A, $N = 10$, for B, $N = 4$, and for C, $N = 5$. We need only calculate 5 probabilities: the probability of 1 or more hits under systems A and B, the probability of 4 or more hits under systems A and B, and the probability of 5 hits under system C. These probabilities can easily be found using the binomial probability mass function (PMF).

Test II

Under test II, we must calculate the probability of a hit when a random date is applied to the square identified for each signal. For system A, outbreak pattern 1, this probability is 0.4, since the signals all occur in square 1 and 4/10 days also have outbreaks in square 1. The probability of 4 or more hits can be found using the binomial PMF with $p = 0.4$, $N = 10$.

For system B outbreak 1, we need two probabilities. First, the probability of one or more hits from the binomial with $p = 0.4$, $N = 1$, for the probability of a hit among signals in cell 1. Here, $p = 0.4$ because the one signal in square 1 must be randomized to any of the 4 days with outbreaks in that square in order to be a hit. We also need the probability of a hit for a binomial with $p = 0.1$, $N = 1$, since the signal in cell 3 could be randomized to day 8 and be declared a hit. The probability of one or more hits from these two is

$0.4(0.9) + 0.6(0.1) + 0.4(0.1) = 0.46$—the probability of a hit in cell 1 and not cell 3, a miss in cell 1 and a hit in cell 3, and a hit in cells 1 and 3.

For system C outbreak 1 we must use conditional reasoning to establish the probability of 5 hits as $(0.4)^4(0.1)$. (The signals in square 1 must be randomized to one of the days on which outbreaks were observed, while the one in square 3 must be randomized to that day.)

For system A outbreak 2, the test II probability is determined from the binomial PMF with $N = 10$, $p = 0.1$, since the signal in square 1 must be randomized to day 1. We use this to find the probability of 1 or more hits. For system B outbreak 2, we can reason directly that 3 of the 4 independent signals (those not in cell 3) each have a 0.1 probability of being assigned to the correct day, while the fourth has a 0.2 probability of being assigned to a day with an outbreak and the probability that all are assigned to a correct day is $(0.1)^3(0.2)$. For system C outbreak 2, three of the signals must be randomized to exactly one day (probability 0.1) while two may be randomized to one of two days (probability 0.2) so that the total probability of exactly 5 hits is $(0.1)^3(0.2)^2$.

For system A outbreak 3, the probability may be calculated as for outbreak 2 since there is again only one outbreak in square 1. For system B outbreak 3, we can most easily figure the probability of no hits is the product of no hits in each cell, or $0.9 \times 0.9 \times 0.9 \times 0.8$, since three cells have a 0.9 probability of no hit and the cell with two outbreaks has a 0.8 probability. The probability of at least 1 hit is therefore $1 - (0.9)^3(0.8)$. For system C outbreak 3, the probability is as for system C outbreak 2, since there is again one outbreak in each of three squares and 2 outbreaks in the remaining square.

Test III

For test III, we must count the number of permutations which generate as many or more hits than were observed. One can think of this as sampling dates for the observed signals without replacement, so that the dates are reassigned to the signals. Fortunately, for systems A and B, it is easy to see that the number of hits that were observed are guaranteed under any permutation. That is, under system A, every day will have a signal in square 1 in every permutation, and under system B, day 1 will have a signal in every square in every permutation. So any signals that were hits in square 1 (system A) or on day 1 (system B) in the observed data will recur in every permutation. This means that the probability of seeing that many hits is exactly 1 for both systems for all outbreaks.

For system C, outbreak 1, there are 5 permutations that are different. One has all five signals occurring when there are outbreaks, and 4 have the signal on day 8 appearing in square 1 (or equivalently the signal in square 3 appearing on day 1, 4, 7 or 9). An alternative reasoning is that there are 5! possible orders of the squares in the signals, and that 4! have 5 hits. In either case, the probability of 5 hits is $1/5 = 4!/5! = 0.2$. (In addition, the probability of 3 hits is 0.8.) Parallel reasoning finds a probability of 0.2 for outbreak 2. For outbreak 3, there are again 5! permutations, but now only two of them have 5 hits; these two are the ones with hits in square 3, with the order of those two reversed. Thus, the probability of 5 hits is $2/5!$, or $1/60 = 0.017$.

REFERENCES

1. Buckeridge D, Burkom H, Moore A, Pavlin J, Cutchis P, Hogan W. Evaluation of syndromic surveillance systems: development of an epidemic simulation model. *Morbidity and Mortality Weekly Report* 2004; **53**(Suppl.):137–143.

2. Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response early aberration reporting system (EARS). *Journal of Urban Health* 2003; **80**(2)(Suppl. 1):i89–i96.

3. Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *BMC Public Health* 2001; **1**(1):9.

4. Lazarus R, Kleinman K, Dashevsky I, Adams C, Kludt P, DeMaria Jr A, Platt R. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerging Infectious Diseases* 2002; **8**(8):753–760.

5. Lober WB, Trigg LJ, Karras BT, Bliss D, Ciliberti J, Stewart L, Duchin JS. Syndromic surveillance using automated collection of computerized discharge diagnoses. *Journal of Urban Health* 2003; **80**(2)(Suppl. 1): i97–i106.

6. Lombardo J, Burkom H, Elbert E, Magruder S, Lewis SH, Loschen W, Sari J, Sniegoski C, Wojcik R, Pavlin J. A systems overview of the electronic surveillance system for the early notification of community-based epidemics (ESSENCE II). *Journal of Urban Health* 2003; **80**(2)(Suppl. 1):i32–i42.

7. Mostashari F, Fine A, Das D, Adams J, Layton M. Use of ambulance dispatch data as an early warning system for communitywide influenza-like illness, New York City. *Journal of Urban Health* 2003; **80**(2)(Suppl. 1): i43–i49.

8. CDC. Framework for evaluating public health surveillance systems for early detection of outbreaks. *Morbidity and Mortality Weekly Report* 2004; **53**(RR-5):1–16.

9. Rogerson P. Monitoring point patterns for the development of space–time clusters. *Journal of the Royal Statistical Society*, *Series A* 2001; **164**:87–96.

10. 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.

11. Kulldorff M. Prospective time periodic geographic disease surveillance using a scan statistic. *Journal of the Royal Statistical Society*, *Series A* 2001; **164**:61–72.

12. Klauber MR. Two-sample randomization tests for space-time clustering. *Biometrics* 1971; **27**:129–142.

13. Dwass M. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 1957; **28**:181–187.

14. Good P. *Permutation Tests*. Springer: New York, 2000; 185–187.

15. Good P. *Permutation Tests* (2nd edn). Springer: New York, 2000.

16. Platt R, Bocchino C, Caldwell B, Harmon R, Kleinman K, Lazarus R, Nelson AF, Nordin JD, Ritzwoller DP. Syndromic surveillance using minimum transfer of identifiable data: the example of the national bioterrorism syndromic surveillance demonstration program. *Journal of Urban Health* 2003; **80**(2)(Suppl. 1):i25–i31.

17. 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.

18. Neises D. *Rate-limiting Factors in Foodborne Outbreak Detection and Non-traditional Foodborne Illness Surveillance in Minnesota*. Division of Epidemiology and Community Health. University of Minnesota: Minneapolis, 2003.

19. 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.