University of Massachusetts Amherst

ScholarWorks@UMass Amherst

Public Health Department Faculty Publication Series

Public Health

2004

A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism

Kenneth Kleinman

Ross Lazarus

Richard Platt

Follow this and additional works at: https://scholarworks.umass.edu/public_health_faculty_pubs

Part of the Epidemiology Commons

A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas, with an Application to Biological Terrorism

Ken Kleinman, Ross Lazarus, and Richard Platt

Since the intentional dissemination of anthrax through the US postal system in the fall of 2001, there has been increased interest in surveillance for detection of biological terrorism. More generally, this could be described as the detection of incident disease clusters. In addition, the advent of affordable and quick geocoding allows for surveillance on a finer spatial scale than has been possible in the past. Surveillance for incident clusters of disease in both time and space is a relatively undeveloped arena of statistical methodology. Surveillance for bioterrorism detection, in particular, raises unique issues with methodological relevance. For example, the bioterrorism agents of greatest concern cause initial symptoms that may be difficult to distinguish from those of naturally occurring disease. In this paper, the authors propose a general approach to evaluating whether observed counts in relatively small areas are larger than would be expected on the basis of a history of naturally occurring disease. They implement the approach using generalized linear mixed models. The approach is illustrated using data on health-care visits (1996–1999) from a large Massachusetts managed care organization/multispecialty practice group in the context of syndromic surveillance for anthrax. The authors argue that there is great value in using the geographic data.

bioterrorism; communicable diseases; epidemiologic methods; generalized linear mixed model; population surveillance; spatial analysis

Recent bioterrorist attacks have highlighted the public health need to detect bioterrorism events as quickly as possible. In the case of anthrax, the initial clinical presentation of persons with acute lower respiratory infection could precede the laboratory diagnosis of anthrax by several days (1). This suggests that surveillance based on rapid reporting of an unusually large number of incident episodes of lower respiratory illness in a community might present an opportunity for early detection of a bioterrorism event. Less dramatically, episodes of small-scale foodborne illness, intentionally caused or otherwise, may currently go unremarked but could be detected through improved surveillance methods.

One way to improve surveillance is to develop methods of identifying temporal and geographic clusters of events that may merit additional evaluation, rather than rely on merely temporal methods. This improvement can be efficient in that data suitable for such analysis are collected routinely by health maintenance organizations and physicians' groups.

The data are increasingly available in a timely fashion as electronic medical records and other real-time data collection systems are adopted. Additionally, geographic information is often available (2). Spatial data are important: Clusters of interest may be localized and of insufficient size to be detected in analyses that consider only an entire region, and/or a lack of a specific location may interfere with treatment and prevention.

Available techniques that can be used with such data include scan statistic (3) and cu-sum (4) approaches, which explicitly aim at surveillance. The referenced articles also offer comprehensive surveys of spatiotemporal surveillance methods. In addition, time-series methods (5) could be employed, and spatial clustering techniques (6) could be adapted to temporal surveillance. As an alternative, we propose a method based on generalized linear mixed models (GLMM) (7).

MATERIALS AND METHODS

Statistical method

The method we propose is based on logistic regression. Consider as a conceptual exercise a logistic regression model predicting caseness on a given day, with predictors describing the day of observation:

$$\text{logit}(\Pr[\text{case}]) = \beta_0 + \beta_1\text{spring} = \beta_2\text{summer} + \beta_3\text{fall} + \beta_4\text{weekday},$$

where "spring," "summer," "fall," and "weekday" are indicator variables describing the season and a nonweekend day. For a given day, we could generate an estimate of the probability that an individual is a case using the estimated β parameters from the above model.

An improvement to this approach would be to incorporate geographic data. This would allow for the possibility that some areas contained populations more likely to become sick, such as older people, or more likely to seek health care, as might be the case for certain cultural groups. We might then modify the conceptual model:

$$\text{logit}(\Pr[\text{case}]) = \beta_0 + \beta_1\text{spring} = \beta_2\text{summer} + \beta_3\text{fall} + \beta_4\text{weekday} + \sum_{i=1}^{I-1}\gamma_1, \qquad (1)$$

where i indexes the areas and each γ_i is the log odds of being a case in area i relative to the Ith of the I areas.

Instead of the ordinary logistic regression discussed conceptually above, we propose the use of GLMM (7) to estimate the probability that each subject under surveillance is a case, in each area, on a given day. (The approach can be used for other summary periods, but we use day as an example.) The proposed model takes the form

$$E(y_{ijt}|b_i) = p_{ijt} \text{ and } \text{logit}(p_{it}) = x_{ijt}\beta + b_i, \qquad (2)$$

where y_{ijt} indicates whether person j in area i is a case on day t, p_{ijt} is the probability that he/she is a case, x_{ijt} is a set of covariates measured for person j and/or area i and/or day t, β is a vector of fixed effects, and b_i is a random effect for area i. One can think of the b_i's as allowing different baseline risks in each area i. The b_i's are typically assumed to have a multivariate normal distribution with a mean of 0.

The estimated b_i's are sometimes called the "shrinkage" estimators, because they are, loosely, a weighted average of the log odds ratio γ_i from model 1 and the mean of these odds ratios. The weights depend on the number of persons in area i, among other things. The fewer persons in area i, the greater the weight of the overall mean; that is, the greater the shrinkage. The shrinkage estimators have a smaller mean squared error than the simple logistic regression estimators for the areas, though they are not unbiased. Shrinkage estimators are sometimes described as "borrowing strength" across the clusters (6). The borrowed strength is why we generally prefer the GLMM to the simple logistic regression.

The covariates x_{ijt} could include 1) measures on the individuals j, such as their age and gender; 2) measures on area i or its population, such as the population density or the median income; or 3) descriptors of day t, such as measures of yearly patterns (be they seasonal indicators or trigonometric functions, etc.), the relative humidity, or the day of the week. Any covariates could vary over time. Finally, the model might incorporate autocorrelation by including recent days' counts as predictors.

We recommend that individual-level covariates be used whenever possible, since their predictive value may be great. In practice, however, their inclusion can make fitting models computationally difficult. In such cases, where individual-level covariates are excluded, the data are effectively a count of the number of cases and noncases in each area i on each day t. A simpler form of the model can then be used, with the notation

$$E(y_{it}|b_i) = n_{it}p_{it} \text{ and } \text{logit}(p_{it}) = x_{it}\beta + b_i, \qquad (3)$$

where

$$y_{it} = \sum_{j=1}^{n_{it}} y_{ijt},$$

n_{it} is the number of individuals in area i on day t, and p_{it} is the probability that any resident of community i is a case on day t. Covariates that differ between members of any community are omitted from x_{ijt}, leading to the simpler x_{it}. Model 3 might be referred to as "binomial logistic regression," as opposed to the Bernoulli logistic regression depicted in model 2, which is commonly referred to simply as logistic regression. The practical effect of this change is wholly computational, assuming that no individual-level covariates are used in either case. The interpretation of the model estimates is identical, but instead of

$$\sum_i n_{it}$$

lines of data for each day, there are only i lines of data for each day.

In most practical applications, including the example presented below, we expect that individual-level covariates will be impractical, and we use the representation shown in model 3 throughout the remainder of the article.

The model is fitted using data from a "historical period" including only days before the surveillance day. Days included might be all available previous days, or certain days could be omitted because of known anomalies in count or conditions such as weather. After fitting of the model, estimated values $(\hat{\beta}, \tilde{b}_i)$ of the fixed effects β and the random effects b_i are used to calculate $\hat{p}_{it}$, the estimated probability of being a case for a given surveillance day t in a given area i:

$$\hat{p}_{it} = \frac{\exp x_{it}\hat{\beta} + \tilde{b}_i}{1 + \exp x_{it}\hat{\beta} + \tilde{b}_i}.$$

Note that day t is specifically excluded from the historical period. This is the surveillance problem: to use the past to evaluate the present.

The probability of seeing z or more cases is calculated from $\hat{p}_{it}$, using the binomial probability mass function:

$$\Pr(\geq z \text{ cases}) = \sum_{s=z}^{\infty} \Pr(Z=s) = 1 - \sum_{k=0}^{z-1} \Pr(Z=k);$$

$$\Pr(Z=k) = \binom{n_{it}}{k}\hat{p}_{it}^k(1-\hat{p}_{it})^{n_{it}-k}.$$

This probability can be thought of as a p value for the null hypothesis that the cases in the area came from a binomial distribution with parameter $\hat{p}_{it}$.

There can be many areas, and the surveillance is repeated daily. Thus, there may be a considerable problem with multiple testing. To avoid this, we suggest not reporting the p value directly. Instead, consider the following transformation: $(p \text{ value} \times NT)^{-1}$ = "expected time," where NT is a relevant number of tests.

The expected time is the number of times NT tests would have to be performed so that one would expect to see exactly one p value as small as or smaller than the one that was observed. For example, if NT = 100, a p value of 0.005 would be expected to appear once in two sets of 100 tests, that is, $(0.005 \times 100)^{-1} = (0.5)^{-1} = 2$. A p value of 0.0001 would be expected to appear once in 100 sets of 100 trials. Note that because of the discrete nature of the binomial distribution, the expected time may be conservative (8). A useful NT would be the number of tests performed in a year; this would allow statements such as "the observed event is expected once every 10 years." One advantage of this approach is that larger values of the expected time are more alarming or unusual, as opposed to the counterintuitive interpretation of smaller p values as more unusual. It also allows the user to focus on the unusualness of the observed data in terms of time, rather than on attempting to evaluate the meaning of a p value in this context.

Example setting and data

We previously described surveillance that uses an automated electronic medical record system among members of a Massachusetts health maintenance organization (2). Briefly, for each health-care encounter, a clinician enters diagnoses, to which *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM), codes are attached. Diagnoses are then grouped into syndromes (2). The diagnosis is available for data analysis immediately. The work was approved by the Harvard Pilgrim Health Care (Boston, Massachusetts) institutional review board.

Patients' addresses are geocoded, which provides the US Census tract of residence. Although the accuracy of commercial geocoding has recently been called into question (9), the geocoding used in this article attached an exact location to 94 percent of addresses. A census tract location was found for an additional 1 percent of subjects. The addresses are used for billing and other business purposes, which suggests that they are sufficiently accurate for surveillance. The data set used here represents the ambulatory medical encounters of approximately 240,000 geocodable individuals—about 10 percent of the population in the region of eastern Massachusetts described below.

A cluster of inhalational anthrax is one of the conditions the system is intended to identify. Typically, inhalational anthrax begins with a nonspecific prodromal phase in which the sufferer may experience fever, dyspnea, cough, and chest discomfort. During this phase, neither physical examination nor any widely used diagnostic test suggests an unusual illness. Diagnosis usually occurs after 2–4 days, when respiratory failure and hemodynamic collapse may ensue. By that time, chest radiographs show an unusual pattern of mediastinal widening (1).

The lower respiratory infection syndrome used in our system was designed to capture cases of anthrax in the prodromal phase. It incorporates 119 ICD-9-CM codes including influenza, pneumonia, bronchitis, and cough. As would be expected, incidence rates are much higher in the winter than in the summer (2). Spatial clusters as well as temporal clusters are expected, because of the contagious nature of illnesses that dominate the visits associated with lower respiratory infection.

The proposed statistical method is intended to alert authorities when local increases in lower respiratory infection occur. This will allow health officials to perform additional evaluation of affected individuals and/or advise clinicians to have heightened suspicion when evaluating other persons with respiratory illnesses. If the local increases reflect prodromal cases of anthrax, the system will help to alert authorities in a timely fashion. The nonanthrax cases of lower respiratory infection are the "noise" against which a "signal" of anthrax-related complaints must be detected. Because of the prodrome, an unusual cluster of cases caused by anthrax might be detected in this way before it was possible to diagnose its cause.
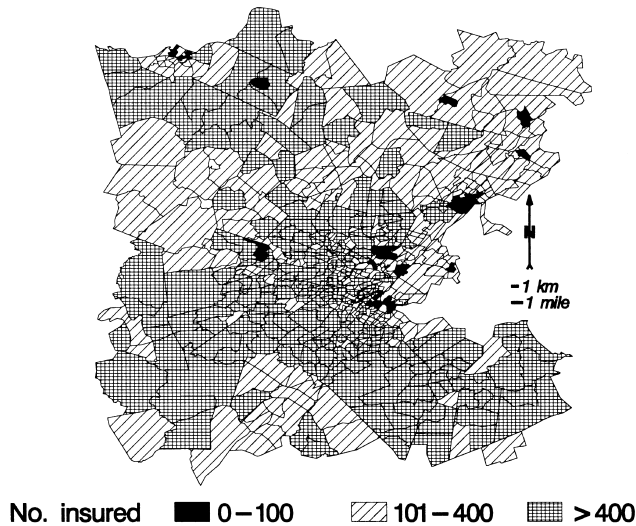
FIGURE 1. Number of subjects covered by a health maintenance organization in each census tract in a region of eastern Massachusetts as of October 1999.

The data set used here incorporates cases of lower respiratory infection syndrome that occurred between January 1, 1996, and October 31, 1999. There were 80,683 lower respiratory infection episodes during this period for which the patient's residence was in one of the 529 census tracts with centroids in the Greater Boston, Massachusetts, area, between west longitudes 70.85 and 71.4 and north latitudes 42.15 and 42.65. A map depicting the number of patients residing in each census tract in October 1999 shows that while a handful of tracts had very few subjects, most had more than 400 (figure 1).

The model we discuss below was generated as an example and is not intended to be exhaustive, parsimonious, or prescriptive. We modeled incident episodes of lower respiratory infection that occurred during the 1,394 days between January 1, 1996, and October 24, 1999. To account for seasonality, we fitted indicator variables for 11 of the calendar months. To adjust for within-week patterns, we included indicators for 6 days of the week. We also included a linear secular time trend to account for any changes in the incidence of disease or in ICD-9-CM coding-choice habits. We did not include any individual-level covariates, since this was impossible computationally. The census tracts are used as the areas discussed in the previous section. We did not include covariates describing the census tracts.

The model was fitted using the GLIMMIX macro (10) in SAS, version 8.2 (11). Note that GLIMMIX uses penalized quasilikelihood estimation, which causes some bias (12); the many repeated values for each census tract should have diminished the bias here. The advantage of this fitting method is that it uses a great deal less computer memory and time than methods that employ numeric integration. This is a meaningful advantage for data sets as large as the one used in this example. (In general, closed-form solutions are not available, so either approximation or numeric integration is necessary.)

We assumed that the random effects for the census tracts were independent. An assumption of spatially correlated random effects could also be appropriate; including spatial correlation would imply that areas close to one another have similar baseline risks of illness. However, a practical problem with incorporating such correlation is that much available software (e.g., S-Plus (13) or Stata (14)) does not allow spatially correlated random effects with longitudinally repeated nonnormal observations. SAS (10, 11) allows it with some modification but could not be employed in this large data set using available hardware. While BUGS (15), which uses Markov chain Monte Carlo methods to fit Bayesian models, might be suitable, this method requires a great deal of care and time, which would preclude the timely repeated fitting of the model necessary for ongoing surveillance. Use of a computationally complex method would also limit applicability because of the specialized knowledge required.

There is also a plausible rationale for avoiding spatial correlation in this example. This region around Boston is made up of many demographically and economically distinct towns and neighborhoods. Towns that border one another may share little besides physical proximity. In this context, smoothing across area may actually not be helpful; there is little reason to think that neighboring areas should have similar baseline risks.

RESULTS

The results of fitting the model are presented in tables 1 and 2. Table 1 shows that the number of cases varies significantly from month to month. For example, the odds of a health-care visit are twice as high in December as in May. There is little difference between December and January, however. Similarly, the odds of a visit are highest on Mondays, nearly three times the odds on Saturday, while the two weekend days have similar odds. There is a statistically significant linearly increasing trend over the 4 years. However, this trend is small, with the relative odds of a visit for lower respiratory infection increasing by only 3.3 percent each year. Not tabulated is the fact that the global likelihood ratio tests for a month effect and a day-of-the-week effect were both statistically significant, with p values less than 0.0001.

The fact that the variance of the random effects, σ_b^2, is significantly greater than 0 implies that the differences between tracts contribute meaningfully to the model. The inclusion of census tract-specific covariates might reduce the variability among the random effects, but it is unlikely that all unique features of a census tract could be accounted for in this fashion. The random census tract effect absorbs the unmodeled unique features of the census tract.

TABLE 1. Parameter estimates and odds ratios for being a case from a model including an intercept, a month, a secular time trend (centered, measured in years), and a random effect for each US Census tract in the study region, eastern Massachusetts, 1996–1999*

	Parameter estimate	95% CI†	Odds ratio	95% CI	p value
Intercept	−8.652	−8.690, −8.615			
Month (relative to December)					
January	0.005	−0.024, 0.034	1.005	0.976, 1.035	0.7732
February	0.076	0.047, 0.106	1.079	1.048, 1.111	<0.0001
March	−0.369	−0.401, −0.337	0.691	0.670, 0.714	<0.0001
April	−0.606	−0.640, −0.571	0.546	0.527, 0.565	<0.0001
May	−0.717	−0.752, −0.682	0.488	0.471, 0.506	<0.0001
June	−0.807	−0.843, −0.771	0.446	0.430, 0.463	<0.0001
July	−1.051	−1.089, −1.012	0.350	0.336, 0.366	<0.0001
August	−1.001	−1.039, −0.963	0.368	0.354, 0.382	<0.0001
September	−0.677	−0.711, −0.642	0.508	0.491, 0.526	<0.0001
October	−0.303	−0.334, −0.271	0.739	0.716, 0.762	<0.0001
November	−0.096	−0.128, −0.063	0.909	0.880, 0.938	<0.0001
Day (relative to Saturday)					
Sunday	−0.022	−0.060, 0.016	0.978	0.942, 1.016	0.2614
Monday	1.087	1.056, 1.118	2.966	2.875, 3.059	<0.0001
Tuesday	0.972	0.941, 1.003	2.644	2.562, 2.278	<0.0001
Wednesday	0.901	0.869, 0.933	2.463	2.386, 2.542	<0.0001
Thursday	0.838	0.806, 0.870	2.312	2.239, 2.388	<0.0001
Friday	0.806	0.774, 0.838	2.238	2.167, 2.312	<0.0001
Year	0.033	0.300, 0.036	1.0335	1.030, 1.037	<0.0001
σ_b	0.033	0.028, 0.038			<0.0001

* The model was based on 250,000 subjects covered by a health maintenance organization in eastern Massachusetts between 1996 and 1999.

† CI, confidence interval.

Table 2 shows 11 census tracts approximately at the deciles of the random effects; we show the exact rank. We chose these 11 census tracts as an illustrative sample of the 525 populated census tracts in the model. In the table, we show the exponentiated random effects (interpreted as the odds ratio in the tract relative to the average tract), the estimated parameter $\hat{p}_{it}$ from the model, the number n_{it} of subjects living in the census tract, the number of visits on October 25, 1999, and the probability of seeing this many or more cases. Note that only data through October 24, 1999, were used in estimating the model. A detailed example of the calculation of $\hat{p}_{it}$ is provided in the Appendix.

Among these 11 census tracts, only one, tract 250173336, has a count that seems unusual, with a p value for two or more visits of 0.0062. This is Pr(≥2 cases) = 1 – Pr(0 cases) – Pr(1 case) = 1 – (0.000633)^{315} – 315 × (0.000367) × (0.999633)^{314}, using the estimated $\hat{p}_{it}$ from table 2. For a day, NT = 525, so the expected time is 0.31 of a day ((0.0062 × NT)$^{-1}$ = 0.31)—that is, more than three times every day. This count is unlikely to merit further attention.

This simple example mimics surveillance: We fit the model using all of the data available up to a given point, esti-

mate the parameters of the binomial distribution for a future point, and evaluate the observed counts. Table 2 shows the extent of the variability between census tracts and the effect that this can have. Three census tracts in the table had two visits, but while the numbers of subjects living in those census tracts were roughly similar, the p values differ by a factor of 3. Of the three, the census tract with the largest random effect has the largest p value, despite being only 80 percent the size of the largest census tract of the three.

We carry the example further in figure 2, where we plot the three census tracts with the smallest p values. (This limit was selected because three census tracts could be plotted comfortably with text.) We note the number of cases seen in these census tracts and the p value associated with that observed count. A p value of 0.0062 was calculated above to have an expected time of 0.31 days. A p value of 0.0036 has an expected time of 0.52 days. A p value of 0.00005 has an expected time of 38 days, meaning that we should expect counts this unusual only once every 38 days if the model is correct. This is unusual enough that a public health department might investigate further by evaluating the individuals

TABLE 2. Eleven US Census tracts, the rank of their random effects, and their $e^{\bar{b}_i}$, $\hat{p}_{it}$, and n_{it} for t = October 25, 1999; the actual number of health-care visits observed on that day; and the probability that this many or more visits would be made in that census tract on that day, eastern Massachusetts, 1996–1999*

Census tract†	Rank	$e^{\bar{b}_i}$	$\hat{p}_{it}$	n_{it}	No. of health-care visits	Probability of that no. of visits or more
173531	1	0.642	0.000247	499	0	1
173689	54	0.812	0.000312	527	1	0.1516
251010	106	0.879	0.000338	2,231	0	1
250814	161	0.920	0.000353	275	0	1
173336	208	0.959	0.000367	315	2	0.0062
173734	261	0.997	0.000383	487	0	1
250820	314	1.038	0.000399	435	1	0.1593
092111	367	1.083	0.000416	186	0	1
214194	420	1.138	0.000437	467	2	0.0182
214042	469	1.220	0.000469	1,250	0	1
092102	523	1.486	0.000571	376	2	0.0200

* The model was based on 250,000 subjects covered by a health maintenance organization in eastern Massachusetts between 1996 and 1999.

† Tract numbers shown omit the state code ("25") and a leading 0 from the county code.

contributing to the unusual count; these persons can be easily identified from health-care-provider records.

We replicated this process for two other days—February 11, 1999, and August 6, 1999—using the data between January 1, 1996, and the previous day to estimate $\hat{p}_{it}$. We chose these months because their odds of lower respiratory infection are relatively high and low, respectively, and these specific days because in each case there was an unusual-seeming count. Smaller counts make the census tracts unusual in the summer than in the winter. For February 11, an observed count of six had a p value of 0.00000022, or an expected time of 8,641 days (23.7 years). For August 6, a count of three had a p value of 0.000014, or an expected time of 135 days.
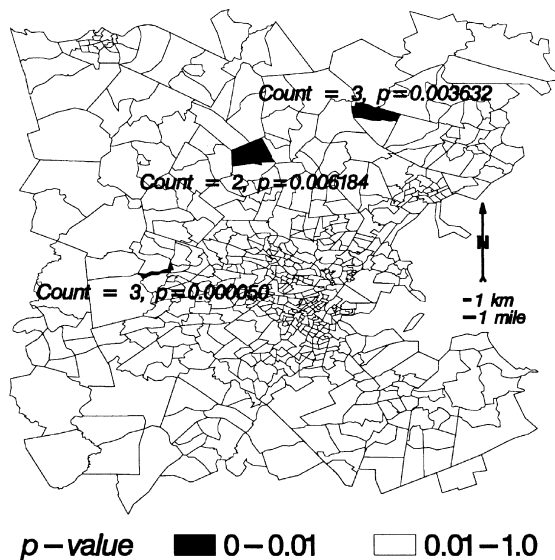
DISCUSSION

The proposed approach can accommodate temporal clustering, secular and seasonal trends, and arbitrarily numerous geographic regions. In addition, it can be fitted using widely available software. Finally, the results of the analysis are adjusted for the size and unique features of the population residing in each region. The results of the model can be presented using a probability-based metric that is adjusted for covariates and unmeasured features, as well as multiple testing. For the data set described here, 3–6 cases in a census tract in one day are often unusual enough to be expected to appear only once in a given month (16).

The logistic GLMM is not the only way to estimate the probability of being a case. An ordinary logistic regression model could be used, with an indicator variable for each area. We recommend the GLMM mainly because the size of the population under surveillance in each area is often variable, and mixed models are often used in such cases to "borrow strength" across the areas (6). A Poisson regression analysis, either using GLMM or with indicators, could also be an appropriate substitute. Finally, this is a relatively crude space-time model. More sophisticated models have appeared (17, 18) and could be adapted to the current surveillance

FIGURE 2. Census tracts with the most unusual counts of lower respiratory illness episodes (p < 0.01), based on the probability mass function derived from the model for October 25, 1999. The model was based on 240,000 subjects covered by a health maintenance organization in eastern Massachusetts between 1996 and 1999.

purpose. However, the implementation of a sophisticated model in a large data set would be costly in time and effort. The relatively large scale of surveillance data sets does make some compromise attractive. This is particularly relevant because of the appeal of expanding real-time cluster detection to include considerably larger populations.

Data on individuals can be included with individual-level covariates, summarized within area, or stratified by some individual-level covariates and then summarized within area. We have alluded to the computational advantage of summarizing the data. Another advantage of summarizing within area is that tables with threshold values can be generated ahead of time, reducing surveillance to a simple look-up task once the model has been fitted. If individual-level covariates were included, the p value and the expected time would depend on which subjects were cases on a given day, making this useful and simple implementation impossible.

The proposed method has advantages and disadvantages relative to alternative techniques. Scan statistic (3) and spatial cu-sum (4) approaches can detect elevated counts over several contiguous areas in a natural fashion, while the proposed technique cannot. On the other hand, these techniques cannot easily incorporate continuous covariates and may be less sensitive to elevated rates in a single area. Time-series methods (5) detect elevated rates across an entire region—a situation the proposed method may lack power against—but would be less sensitive to a smaller number of spatially focused cases, a particular strength of the proposed method.

There are several important limitations of the proposed statistical approach. Problems associated with any division of space are well known in cluster-identification research. For instance, a cluster divided between two areas may fail to attract attention (19). A similar question is how to choose among available spatial divisions. A conservatism of the "expected time" procedure is that it assumes that all tests are independent. This would be troubling if spatially correlated random effects were found useful; there would then be a discrepancy between the model and the evaluation method.

No analysis of a data set of this type is likely to detect an occurrence like the 2001 US bioterrorism attack using anthrax: There were too few cases, and they were more connected by work site than by home address. Nonetheless, it is easy to imagine an attack that could be detected in this kind of data set: exposure through the ventilation system of a restaurant, for example. In general, the data collection system described has an advantage in detecting ailments with nonspecific prodromes, such as anthrax, or nonspecific symptoms more generally. Arguably, this includes anthrax, botulism, plague, smallpox, and tularemia—all of the Centers for Disease Control and Prevention category A bioterrorism agents except viral hemorrhagic fever (20).

There are other limitations of the type of data considered in the example. The nonrepresentativeness of the population under surveillance raises questions about its suitability for bioterrorism surveillance. This concern is heightened by the fact that 5 percent of addresses in the data set could not be geocoded. Persons subject to surveillance under this system are generally better educated, wealthier, and more likely to be White than the general public. To the extent that these characteristics would protect them from exposure or decrease the chances that they would seek health care, the example data set would be less likely to reflect bioterrorism than a representative population. However, while the sample is not representative, it is quite large, which would help officials avoid completely missing a bioterrorist attack just by chance.

Additional work in statistical models for cluster surveillance will be required to resolve some of the limitations described here. One important practical development will be to assess adjacent areas with p values that are small but not small enough to suggest further investigations by themselves. A formal preliminary step for determination of the best spatial areas might be developed. The impact on sensitivity of varying proportions under surveillance in different areas should be assessed. Finally, as computing speed increases and coding efficiency improves, it may become practical to fit models that consider each individual, rather than summing data within an area.

8

REFERENCES

1. Recognition of illness associated with the intentional release of a biologic agent. MMWR Morb Mortal Wkly Rep 2001;50:893–7.
2. Lazarus R, Kleinman KP, Dashevsky I, et al. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health 2001;1:9.
3. Kulldorff M. Prospective time periodic geographic disease surveillance using a scan statistic. J R Stat Soc A 2001;164:61–72.
4. Rogerson PA. Monitoring point patterns for the development of space-time clusters. J R Stat Soc A 2001;164:87–96.
5. Chatfield C. Time-series forecasting. New York, NY: Chapman and Hall, Inc, 2000.
6. Lawson AB. Statistical methods in spatial epidemiology. New York, NY: John Wiley and Sons, Inc, 2001.
7. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc 1993;88:9–25.
8. Waller LA, Turnbull BW, Clark LC, et al. Spatial pattern analyses to detect rare disease clusters. In: Lange N, Ryan L, Billiard L, et al, eds. Case studies in biometry. New York, NY: John Wiley and Sons, Inc, 1994.
9. Kreiger N, Waterman P, Lemieux K, et al. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 2001;91:1114–16.
10. SAS Institute, Inc. GLIMMIX: a SAS macro for fitting generalized linear mixed models using Proc Mixed and the Output Delivery System (ODS). Cary, NC: SAS Institute, Inc, 2003. (Available from: http://ewe3.sas.com/techsup/download/stat/glmm800.sas).
11. SAS Institute, Inc. SAS OnlineDoc, version 8.2. Cary, NC: SAS Institute, Inc, 2001.
12. McCulloch CE, Searle SR. Generalized, linear, and mixed models. New York, NY: John Wiley and Sons, Inc, 2001:281–4.
13. Insightful Corporation. S-Plus 6 for Windows user's guide. Seattle, WA: Insightful Corporation, 2001.
14. Stata Corporation. Stata statistical software, release 8.0. College Station, TX: Stata Corporation, 2003.
15. Medical Research Council. The BUGS Project: WinBUGS. Cambridge, United Kingdom: MRC Biostatistics Unit, University of Cambridge Institute of Public Health, 2003. (World Wide Web URL: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml).
16. Lazarus R, Kleinman K, Dashevsky I. Use of automated ambulatory-case encounter records for detection of acute illness clusters, including potential bioterrorism events. Emerg Infect Dis 2002;8:753–60.
17. Waller LA, Carlin BP, Xia H, et al. Hierarchical spatio-temporal mapping of disease rates. J Am Stat Assoc 1997;92:607–17.
18. Bernardinelli L, Clayton D, Pascutto C, et al. Bayesian analysis of space-time variation in disease risk. Stat Med 1995;14:2433–43.
19. Waller LA. Statistical power and design of focused clustering studies. Stat Med 1999;15:765–82.
20. Centers for Disease Control and Prevention. Emergency preparedness and response. Biological agents/diseases. Atlanta, GA: Centers for Disease Control and Prevention, 2003. (World Wide Web URL: http://www.bt.cdc.gov/agent/agentlist-category.asp).

APPENDIX

Here we show examples of the calculation of $\hat{p}_{it}$. Suppose we were planning surveillance for a Tuesday in January 2005, in census tract 250214194, using effect estimates from tables 1 and 2. The estimated logit would be –8.652(intercept) + 0.005 (January effect) + 0.9722 (Tuesday effect) + 0.165 (secular time trend) + 0.129 (tract effect) = –7.3808. The $\hat{p}_{it}$ for that day in that tract would be $e^{-7.3808}/(1 + e^{-7.3808}) = 0.0006$, meaning that the estimated probability that a person in that tract will be a case on that day is 0.0006. Alternatively, consider a Sunday in July 2004. The estimated logit would be $-8.652 - 1.051 - 0.022 + 0.149 + 0.129 = -9.447$. The $\hat{p}_{it}$ would be 0.00008, meaning that in that tract, the probability of being a case is approximately 7.5 times greater on a Tuesday in January than on a Sunday in July. In an actual application, we would expect to use more recent data to estimate model parameters.