



## Invited Commentary: Surveilling Surveillance—Some Statistical Comments

Lance A. Waller

From the Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA.

Received for publication August 11, 2003; accepted for publication September 15, 2003.

I congratulate Kleinman et al. (1) on their thoughtful application of generalized linear mixed models (GLMM) to disease surveillance in space and time. In this commentary, I amplify some appealing features of the approach, provide an overview of data issues that may affect the field performance of a surveillance system based on such a method, and discuss several technical issues.

### Attractive features of the approach

The authors' approach offers a movement from statistical testing to statistical modeling for disease surveillance. Traditionally, statistical methods for surveillance tend to evolve from a hypothesis-testing framework, wherein one "detects" an outbreak (anomaly, cluster, etc.) as a "statistically significant" departure from a null hypothesis defined as the lack of an outbreak (e.g., constant age-specific incidence proportions or monthly seasonal incidence proportions based on historical data). The current approach uses GLMM to provide predictions of the expected number of cases under the model in the absence of an outbreak and then to compare observed case counts with those model-based expected values. On one level, the goals appear the same, but the end result of a hypothesis-testing approach tends to be an assessment of statistical significance (e.g., a  $p$  value) reflecting whether or not we have sufficient evidence to reject our null hypothesis, while the end result of the model-based approach is a description of which data appear to deviate from the model and by how much. As a result, a testing approach tends to focus on a "yes/no" ("detect"/"nondetect") assessment, but basic features of the surveillance problem considerably complicate formal statistical inference in this setting. For instance, the ongoing temporal nature of surveillance is somewhat similar to methods of sequential analysis, but without an endpoint, and somewhat similar to methods in statistical quality control, but conducted for multiple regions and/or outcomes simultaneously. A model-based approach does not solve these problems per se but places emphasis on the description of patterns rather than solely on assessing their significance via a binary decision. Modeling also offers

the opportunity to improve the model (through inclusion of additional covariates, etc.) in order to better characterize and understand patterns in the data, rather than reach a simple "significant"/"nonsignificant" conclusion. While admittedly glossing over a myriad of technical details regarding the mathematical subtleties of statistical testing and modeling, I personally find the model-based approach better suited to the exploration and understanding of observed patterns.

The proposed GLMM approach has several appealing features. First, it allows ready incorporation of covariate effects to permit adjustment for regional and temporal variations in known risk factors (e.g., age and sex). In addition, the random effects provide a powerful tool for incorporating possible correlations within and between small geographic areas and/or time periods. The "shrinkage" aspect of GLMM estimation, as described by the authors, coupled with the specification of random effects, allows the model to "borrow strength" to improve precision where it is needed most (i.e., the approach borrows more "outside" information for the least precise crude estimates). Very generally, one can consider the specification of random effects to define "from whom each estimate should borrow information." That is, the random effects define which observations have similarities and correlations unaccounted for by the "fixed-effect" covariates. The authors define random intercept terms for each small region, but one could also consider random intercepts for each neighborhood (to allow within-neighborhood correlations due to unmeasured behavioral similarities in residents of each city neighborhood) or spatially correlated random effects (to permit spatial correlation between regions, allowing for broad spatial trends unaccounted for by the covariates within the model). While random effects offer a broad set of possibilities, not all of these possibilities are easily fitted with current software. In the spatial setting, users often build from the work of Clayton and Kaldor (2) and Besag et al. (3) and use Markov chain Monte Carlo algorithms to fit GLMM with spatial random effects. As Kleinman et al. noted (1), such methods (while increasing in popularity) currently do not provide the sort of quick and repetitive kinds of analyses (analyze today's data by

tomorrow) necessary for automated application in the surveillance setting. For the interested reader, a paper by Agresti et al. (4) provides an excellent overview of the fitting and application of GLMM.

### Data features affecting field performance

While the authors' method is quite general and offers opportunities for interesting analyses, there are also some features of surveillance data that might affect how accurately the approach can assess an incident event of interest. In a word, "heterogeneity" summarizes the comments below. In surveillance, we are seeking a "signal" in very noisy data, and the noise may induce patterns that may or may not correspond to the patterns of concern.

First, the timing of illness reports is not discussed but may prove important, particularly if there are heterogeneities between reporting units—for example, if plan members are seen at medical practices in different parts of the city. For data such as those discussed here, medical practices are motivated to submit claims in a timely manner, but this alone may not guarantee the same time-to-report for each incident case. In the analysis of disease registries, one often considers a time lag to account for delays in reporting, but this option is less desirable in the present application.

Next, geocoding accuracy is important, as Kleinman et al. noted (1). However, the point is worth reiterating. In particular, a 90 percent match rate means that 90 percent of the addresses were assigned a fixed location on the map, but this may not mean that 90 percent of the addresses were assigned to their corresponding correct location. Some geocoding algorithms assign addresses evenly between street corners, while others are based on more accurate cadastral maps based on taxed parcels of land. In addition, a given address may be accurate enough for mail delivery (including medical bills) but not accurate enough for assignment of a residence to the proper side of the street, which can affect accurate assignment to the correct census block, block group, or census tract. The referenced work by Kreiger et al. (5) provides an introduction to the fascinating issue of geocoding in the realm of public health.

Third, the authors note that approximately 10 percent of the study population is included in the membership database. They note that this subpopulation differs slightly from the population at large in sociodemographic terms, but this subpopulation may also differ geographically. If the proportion of plan members in each census tract is the same for all tracts, the membership "filter" simply reduces the number of incident events in each small region without affecting the overall pattern. However, if the proportion of members varies appreciably from tract to tract, this filter may yield a spatial incidence pattern that differs from the true underlying and unobserved pattern and may affect the ability of the approach to detect anomalies in some portions of the study area.

In addition, if we couple the spatial "filter" of membership with the heterogeneous population sizes across tracts, we have spatially varying sensitivity and specificity to detect events of various sizes in different locations. The overall impact of spatially varying performance of cluster detection

techniques provides a context for any application and merits detailed consideration, since the probability of detecting a cluster of a given effect size and geographic extent varies from location to location.

Another issue complicating surveillance is population migration. While managed care systems may be motivated to keep up-to-date records of members' addresses, capturing residential migration, most people engage in daily migration from home to work, often with many stops in between (e.g., a grocery store, train station, restaurant, or day-care center). If the relevant location of exposure is unrelated to residence location (e.g., a clandestine biochemical release in a busy transfer station for a commuter rail system), an accompanying detectable aggregation of residence locations may not occur. Similarly, the suggested example of a restaurant-based foodborne outbreak also may not result in a spatial cluster if the restaurant patrons came from several different neighborhoods.

Finally, Kleinman et al. use the anthrax episode of 2001 to illustrate the need for public health surveillance, but we note that the particular details of that event would be unlikely to generate spatial and/or temporal clusters detectable by the proposed approach. Specifically, the contaminated letters sent to different (work) locations at somewhat different times would be unlikely to generate the types of health-outcome patterns that would trigger the proposed system. This example and those above illustrate that we are unlikely to obtain a "silver bullet" surveillance technique and that all approaches will differ with respect to the kinds of events they can and cannot detect.

### A few statistical details

Kleinman et al. note that the logistic regression model could be based on individual-level data if relevant outcome and covariate data were available. Mathematically this is true, but this extension merits comment, for two reasons. First, increasing interest in confidentiality will probably have an impact on the availability of individual-level data and on the structure used for reporting results. Second, moving from individual data to aggregate data introduces an ecologic aspect to the analysis, so estimates of odds ratios may not correspond precisely between individual and aggregate data. Wakefield (6) offers a thoughtful discussion of ecologic analysis and the role of spatial effects in ecologic analysis. This said, we should recall that the goal of the authors' surveillance approach is accurate prediction rather than accurate estimation of particular risk factors, and the ecologic nature of the analysis may be of secondary importance.

The authors provide an interesting interpretation for each unusual cluster, namely the time expected to observe as extreme a result, under the model. They note that the proposed estimate of this time (based on a Bonferroni-type combination of independent tests) may be conservative because of correlations among results for the regions. A second factor affecting the accuracy of the measure is the discreteness of the data in each region (which varies along with the population size in each region). As Besag and Newell (7) and Waller et al. (8) have discussed, for rare outcomes, the observable  $p$  values fall into a discrete set of

values. That is, for discrete distributions, there is only a finite number of  $p$  values one can obtain, and for very rare events we might find that the addition of a single case reduces the local  $p$  value from 0.051 to, say, 0.0001. Based on this idea, the “expected time” is more accurately estimated via summing the probability that each region exceeds the observed  $p$  value of interest over the number of relevant tests. That is, if we observe a small  $p$  value in region  $i$ , such as 0.00005, we must calculate the probability of observing a  $p$  value that small or smaller in each of the other regions, under the model, and then sum these probabilities. This sum is equivalent to the authors’ “expected time” if each of the probabilities is equal to the  $p$  value of interest. This sum provides the correct value when we have continuous measures for each region, but it may differ for discrete data, particularly for surveillance of relatively rare events.

### Conclusion

Kleinman et al. (1) have provided an interesting and valuable contribution to the public health surveillance literature. The proposed model offers the promise of interesting applications and methodological research for some time to come.

### REFERENCES

1. 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. *Am J Epidemiol* 2004;159:217–24.
2. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risk for use in disease mapping. *Biometrics* 1987;43:671–81.
3. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann Inst Stat Math* 1991;43:1–59.
4. Agresti A, Booth JG, Hobert JP, et al. Random-effects modeling of categorical response data. *Sociol Methodol* 2000;30:27–80.
5. 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.
6. Wakefield J. Sensitivity analyses for ecological regression. *Biometrics* 2002;59:9–17.
7. Besag J, Newell J. The detection of clusters in rare diseases. *J R Stat Soc Ser A* 1991;154:327–33.
8. Waller LA, Turnbull BW, Clark LC, et al. Spatial pattern analysis to detect rare disease clusters. In: Lange N, Ryan L, Billard L, et al, eds. *Case studies in biometry*. New York, NY: John Wiley and Sons, Inc, 1994:3–23.