Clustering under the null: one reason for too many signals

Ken Kleinman
Department of Ambulatory Care and Prevention,
Harvard Medical School and Harvard Pilgrim Health Care

OBJECTIVE

I examine the nature and expression of the null hypothesis often used in spatial surveillance. I also show an example of how incorrect specification of the null can lead to excess signals without interesting outbreaks, and argue that this may be a cause of excess signals when using spatial surveillance in public health applications.

BACKGROUND

The tendency of syndromic surveillance systems to generate many outbreak signals is well known. A recent review noted a general state of "relatively low specificity and positive predictive value, with a considerable burden of false alarms." [1] Reasons excess signals may occur include differentially missing or delayed data and poor control for predictable noise. Other problems may be specific to particular cluster detection algorithms. For example, spatial cluster detection requires more complex specification in return for its promise of greater sensitivity and general utility.

In many popular cluster detection methods, including SaTScan™ software [2], the null hypothesis is that the probability of a case appearing in a discrete region (e.g. zip code day) is proportional to the population in that region, effectively a binomial distribution within each region, where all regions share a binomial probability *p*. This is often expressed in English as "complete spatial randomness". The alternative is that some region(s) have a different *p*.

Unfortunately, in many public health applications, uninteresting patterns do not conform to this technical expression of the null. For example, members of the same household or apartment block may become cases at similar times for trivial reasons. This will appear to be a cluster, but may not be of public health significance. In fact, if this happens often enough, the null hypothesis will not be violated in the long run of surveillance, but many signals will be produced. In these situations, the desired null hypothesis of "nothing interesting is going on" is not well represented by the null of equal *p*.

METHODS

We evaluated the performance of Poisson [2] and space-time permutation [3] scan statistics when the data are distributed as under the null of equal *p* and from an alternative null where several regions each day have a larger *p*, at random. We generated data for each of 501 zip codes in eastern Massachusetts under 1) a binomial distribution with N = 23, p=.25 (the usual null) and 2) a mixture of binomials with a .99 probability a binomial with N=100, p = .05 and a .01 probability a binomial with N = 40, p = .5 (the alternative null). For the Poisson scan we used a population of 25,000 for each zip code. Data were generated for 478 days, with the last 382 used for evaluation.

RESULTS

For simulated binomial data with a single *p*, neither the space-time permutation scan nor the Poisson space-time scan generated any signals with a p-value below .05. This is a surprisingly small number of small p-values; we should expect to see about 18 p-values smaller than .05 with this many tests. In contrast, the data generated from the mixture resulted in a p-value smaller than .05 on all but one of 386 days using both the space-time permutation and the Poisson scan.

CONCLUSIONS

The practical null hypothesis of "nothing interesting is going on" may not be well-represented by assuming the regions all have the same probability of cases. If the data are in fact distributed as a mixture of binomials, that is, with excess cases in some regions with no special spatial pattern over time, excess signals may result. This statistical description corresponds to small clusters occurring within, perhaps, families or buildings. The example shows the potential effects of this, albeit in an exaggerated form. This kind of mixture-type null is plausible and may explain to some degree the excess signals produced by spatial surveillance.

In public health applications, spatial analysis is often implemented through SaTScan™ software. [2] Testing in SaTScan™ is based on Monte Carlo reassignment of the cases; the way cases are reassigned represents the null hypothesis. This makes it relatively simple to change the null, compared with theory-based tests. The desired null hypothesis could be rephrased as "there are no clusters bigger than the usual", and this null could be implemented by using a Monte Carlo case reassignment which resulted in a typical number of clusters of typical size. We suggest two permutation strategies for accomplishing this type of Monte Carlo case reassignment.

REFERENCES

[1] Hope K, Durrheim DN, d'Espaignet ET, Dalton C. Syndromic surveillance: is it a useful tool for local outbreak detection? *Journal of Epidemiology and Community Health* 2006; 60:374

[2] Kulldorff M. and Information Services, Inc. SaTScan™ v6.1:Software for the spatial and space-time scan statistics. http://www.satscan.org 2006

[3] Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine* 2005; 2:216-224.