Privacy Protection versus Cluster Detection in Spatial Epidemiology

Karen L. Olson, PhD[1], Shaun J. Grannis, MD, MS[2], Kenneth D. Mandl, MD, MPH[1]

1. *Children's Hospital Boston and Harvard Medical School, Boston, Massachusetts*

2. *Regenstrief Institute and Indiana University School of Medicine, Indianapolis, Indiana*

OBJECTIVE

To investigate the impact of address precision (exact latitude and longitude versus the center points of zip codes) on spatial cluster detection.

BACKGROUND

With the widespread deployment of near real time population health monitoring, there is increasing focus on spatial cluster detection for identifying disease outbreaks. These spatial epidemiologic methods rely on knowledge of patient location to detect unusual clusters. In hospital administrative data, patient location is collected as home address but use of this precise location raises privacy concerns. Regional locations, such as center points of zip codes, have been deployed in many existing systems. However, this practice could distort the geographic properties of the raw data and affect subsequent spatial analyses. The impact of location error due to centroid assignment on the statistical analyses underlying these systems requires study.

METHODS

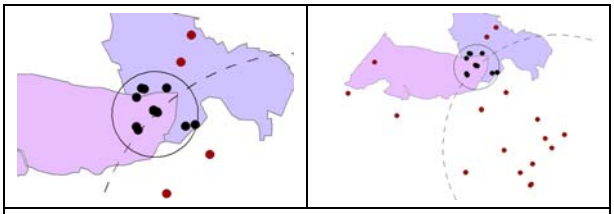
Disease outbreaks were simulated by adding spatially clustered, extra encounters to authentic emergency department (ED) syndromic surveillance data from an urban, pediatric, tertiary care hospital. Baseline data were all encounters with a chief complaint or diagnosis indicative of respiratory illness over 3 years for patients living within 80 km of the hospital.

Simulated clusters were designed to mimic an event of interest: an early signal of an outbreak that first appears as a small geographic cluster. Each cluster contained 10 points and was located along the edge of a circle with a radius of 5 km centered at the hospital. Clusters varied on radius size, from 0.5 to 3 km. Cluster points fell into a single zip code, or they crossed administrative boundaries and were distributed as evenly as possible among 2-4 zip codes.

A spatial scan statistic, as implemented in SaTScan™ (1), was used to detect spatial clustering. A total of 35,200 analyses were performed to test the effect of moving a point from its exact location to the center of a zip code region.

RESULTS

Exact coordinates yielded more (12,858/17,600, 73%) significant clusters than zip code centroids (7,876/17,600, 45%, odds ratio=3.35, 95% CI 3.2-3.5). To assess accuracy, a requirement that significant clusters contain at least half of the original simulated points was imposed. Both a larger absolute number and a larger proportion of the significant clusters met this requirement when exact coordinates were analyzed. 12,016/12,858 (93%) of the significant clusters contained 5-10 simulated points when analyzed as exact coordinates, compared to 6,842/7,876 (87%) of those analyzed as zip code centroids (odds ratio=2.16, 95% CI 1.96 - 2.37). To further assess accuracy, the number of points from the background ED data that were drawn into the significant clusters was counted. Analyzed as exact locations, significant clusters contained fewer additional ED visit points (i.e., points that were not part of the original simulated cluster) (mean=4, std= 10, range 0-111) than when analyzed as zip code centroids (mean=10, std= 21, range 0-157).



Sample result. When points were analyzed at their exact locations (left), the detected cluster contained all of the simulated points (black dots in small circle) and few additional points from the ED visits (red dots outside small circle). However, when the same data were analyzed with points at zip code centroids (right), many more ED visits were included in the detected cluster.

CONCLUSIONS

The spatial cluster detection algorithm performed better when addresses were analyzed as exact locations, particularly when the clustered points crossed administrative boundaries. Use of precise addresses offers improved performance, but must be weighed against privacy concerns when establishing public health data exchange policies.

REFERENCES

1. Kulldorff M, Information Management Services. SaTScan™ version 5.0: Software for the spatial and space-time scan statistics. 2004.

For further information, contact Karen L. Olson karen.olson@childrens.harvard.edu