The cost of obfuscation when reporting locations of cases in syndromic surveillance systems

Caroline Jeffery MA, Al Ozonoff[1] PhD, Laura Forsberg SM , Miriam Nuño PhD and Marcello Pagano PhD

Departments of Biostatistics, Harvard School of Public Health and [1]Boston University School of Public Health

OBJECTIVE

This paper quantifies the effect of not providing full information about the location of patients when dealing with spatio-temporal systems in syndromic surveillance. The study investigates the loss of power to detect clusters when aggregation takes place.

BACKGROUND

In a classical surveillance system one looks for disturbances in the number of cases, but in a spatio-temporal system, not only the number of cases observed but also where they are located is reported. What location is reported, and to which degree of accuracy it is reported are important. At one extreme lies near-perfect information about each case, as with contact tracing; at the other extreme we have no information about location; viz. just that the patient exists, or a temporal system. From maximum spatial precision to no spatial precision, one gains in speed of reporting and privacy; but one loses power to detect outbreaks. For example, in [2] we see that more than one address is better than just a single one. This general point is intuitively appealing, and can be demonstrated.

METHODS

Our population of patients is drawn from three community hospitals serving Cape Cod, Massachusetts ([3]). We simulate a daily reporting of patients' location from this population, and superimpose one of four simulated clusters. Outbreaks are simulated by depositing a cluster atop daily patient volume (simulated by random sampling) and the power to detect these clusters is calculated using the M statistic [4]. This procedure is repeated, using grid points to emulate varying degrees of spatial precision. The coarseness of the grid is varied and the power is then calculated and reported as a function of the surface area of a typical quadrant.

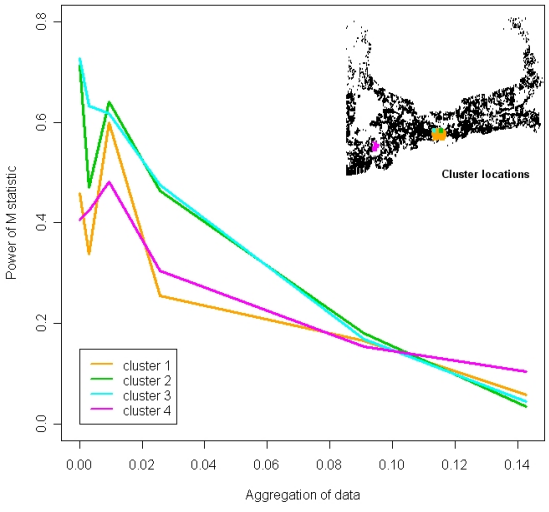

Figure 1 – Power curves for the detection of each of the four clusters in the Cape population, as a function of coarseness of the reporting of the locations of the patients. The quantity on the horizontal is proportional to the surface area of each quadrant in the discretization.

RESULTS

As expected, the level of aggregation is commensurate with loss of power. There is a general downward trend for each of the four simulated clusters, with the power to detect the cluster decreasing as the regions of aggregation grow large (Figure 1).

CONCLUSIONS

The power to detect an outbreak does indeed go down with the aggregation of the data; the greater the loss in spatial precision, the greater the loss of power. Reporting addresses at a level that assures anonymity will result in a loss of power and one should be aware that this loss can be sizeable. If one wishes to maximize the chances of detecting an outbreak, then possibly one should seek alternatives that ensure patients' privacy, and at the same time provide as much spatial information as possible about them.

REFERENCES

[1] M. Meselson, J. Guillemin, M. Hugh-Jones, A. Langmuir, I. Popova, A. Shelokov, and O. Yampolskaya, The Sverdlovsk Anthrax Outbreak of 1979, *Science*, Volume 266, Issue 5188 (November 18, 1994), pp. 1202-1208.

[2] A.Ozonoff, M. Bonetti, L. Forsberg, C. Jeffery and M. Pagano, The distribution of interpoint distances, cluster detection, and syndromic surveillance. Proceedings of the American Statistical Association, Biometrics Section [CD-ROM]. ASA (2004).

[3] A Ozonoff, M Bonetti, L Forsberg, M Pagano. The use of multiple addresses to enhance cluster detection. 2003 Proceedings of the American Statistical Association, Biometrics Section [CD-ROM]. ASA (2003).

[4] Bonetti M. and Pagano M. The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. Stat Med. 2005 Mar 15;24(5):753-73.

The research in this paper was funded in part by a grant from the National Institutes of Health, RO1-AI28076, the National Library of Medicine, RO1-LM007677, and the AIDS Training Grant T32-AI007358. We would like to thank the Institute for Health Metrics for providing us with their data (http://www.healthmetrics.org).

Further Information: M. Pagano, pagano@biostat.harvard.edu