PERGAMON

Health & Place 9 (2003) 273–277

HEALTH & PLACE

www.elsevier.com/locate/healthplace

Visualization of the spatial scan statistic using nested circles

Francis P. Boscoe[a], Colleen McLaughlin[b], Maria J. Schymura[b,*], Christine L. Kielb[c]

[a] *SEER Program, National Cancer Institute, Bethesda, Maryland, MD, USA*

[b] *New York State Department of Health, New York State Cancer Registry, Corning Tower Room 536, Empire State Plaza, Albany, NY 12237, USA*

[c] *Department of Health, Center for Environmental Health, Troy, NY, USA*

Abstract

We propose a technique for the display of results of Kulldorff's spatial scan statistic and related cluster detection methods that provides a greater degree of informational content. By simultaneously considering likelihood ratio and relative risk, it is possible to identify focused sub-clusters of higher (or lower) relative risk among broader regional excesses or deficits. The result is a map with a nested or contoured appearance. Here the technique is applied to prostate cancer mortality data in counties within the contiguous United States during the period 1970–1994. The resulting map shows both broad and localized patterns of excess and deficit, which complements a choropleth map of the same data.

Keywords: Cluster detection; Spatial scan statistics; Prostate cancer; Visualization

Introduction

The choropleth (value-by-area) map is the most common method of depicting disease rates geographically (Walter and Birnie, 1991; Le et al., 1995; Pickle et al., 1996; Devesa et al., 1999). One problem associated with this method is that extreme values are most commonly found in sparsely populated areas, and these areas account for a disproportionate amount of the total map area. Because of this, novice map users are often erroneously drawn to what appear to be areas of high disease rates in large, sparsely populated states such as Wyoming or Nevada, while experienced users may be just as likely to discount a statistically important trend in such areas. This situation underscores the importance of incorporating statistical information into choropleth maps so that genuine excesses can be distinguished from random noise (Pickle et al., 1996; MacEachren et al., 1998; Kulldorff, 1999a).

There are a variety of general cluster detection methods that can help to make this distinction, by defining contiguous areas of excess or deficit that are not likely to have arisen by chance (Openshaw et al., 1987; Turnbull et al., 1990; Besag and Newell, 1991; Kulldorff and Nagarwalla, 1995; Fotheringham and Zhan, 1996; Kulldorff, 1997). These methods involve the generation and comparison of large numbers of circular areas (other shapes are theoretically possible), and tend to identify many similar, closely overlapping circles. Some researchers have opted to display all of these circles, with their density suggestive of the areas with the greatest statistical power (Openshaw et al., 1987; Turnbull et al., 1990; Fotheringham and Zhan, 1996; Timander and McLafferty, 1998). This apparent relationship, however, is biased by the underlying geography of the areas involved. A cluster in Manhattan, for example, would tend to contain more circles, and thus be more visually prominent, than a statistically identical cluster in rural upstate New York, simply as a result of their differing population densities.

Those following Kulldorff's approach, in contrast, have typically reported only the circles with the highest

*Corresponding author. Tel.: +1-518-474-2255; fax: +1-518-473-6789.

E-mail address: mjs08@health.state.ny.us (M.J. Schymura).

1353-8292/03/$ - see front matter
doi:10.1016/S1353-8292(02)00060-6

likelihood ratio that are non-overlapping, provided these circles meet some threshold for statistical significance (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997; Jemal et al., 2000; Gregorio et al., 2001; Sankoh et al., 2001; Forand et al., 2002). While this approach reduces the graphic complexity considerably, it tends to identify large areas with large populations but small elevations in risk, since such areas have the highest statistical power. Smaller clusters contained within these areas that have higher elevations in risk but lower, though statistically significant, likelihood ratios, are ignored. Such sub-clusters might be expected, for example, in the case of a dose–response relationship from an environmental point source, or may arise for many other reasons. In any case, the knowledge of these sub-clusters would certainly contribute to the generation of possible explanatory hypotheses. There have been several examples where sub-clusters have been reported; these have depended on sequential limitation of the maximum allowable circle size (Kulldorff et al., 1997) and an isotonic regression function (Kulldorff, 1999b). Neither of these approaches have been operationalized in the freeware SaTScan computer program employing Kulldorff's method (Kulldorff et al., 1998; Kulldorff and Information Management Systems Inc., 2002).

We use the results of Kulldorff's spatial scan statistic to propose a method of selecting clusters for display that is straightforward and as least as informative as any of the above approaches. While our approach is applicable to any of the general cluster detection methods cited above, we chose Kulldorff's spatial scan statistic because it confers advantages not shared by all of the other methods: it identifies circular clusters of any size, located anywhere within a study area, while controlling for multiple hypothesis testing; it is conceptually straightforward; and is emerging as a widespread tool for identifying areas of unusual disease patterns. Our approach involves stratifying the set of statistically significant circles by relative risk. Within each relative risk stratum, the nonoverlapping circles with highest likelihood are displayed, creating a nested or contoured effect.

Methods

In our analysis, we used 522,994 observed and expected prostate cancer deaths among white males from 1970–1989 in 3053 counties in the contiguous 48 states. These data are available through the National Cancer Institute Cancer Mortality Maps and Graphs web site (National Cancer Institute, 2001). For the purposes of calculating distances between counties and defining groupings of counties, county geographic centroids were used. A grid of over 18,000 points was constructed, consisting of each county centroid and the equidistant points between all combinations of adjacent county centroids. That is, a grid point was added for every instance where two counties adjoin, where three counties adjoin, and where four counties adjoin. The grid design is such to ensure that virtually all possible groupings of between two and four adjacent counties were considered in the analysis, and is more computationally efficient than a regularly spaced grid. For each grid point, the spherical distance to each of the 3053 centroids was calculated and sorted. The sorted lists were used to define groupings of counties with a cumulative number of observed and expected prostate cancer deaths. As the number of counties increases, the shape of the groupings more closely approximates a circle, and following convention we refer to these groupings as "circles". An upper limit was imposed such that no circle could contain more than half of the total prostate cancer deaths. The total number of circles was therefore approximately 27.5 million (18,000 times 3053 divided by 2), which includes some redundant circles.

For each circle, the relative risk and likelihood ratio were calculated based on the cumulative observed and expected deaths contained within it. Relative risk is a measure of the increased or decreased risk associated with being in a particular circle relative to the nation as a whole and simply the ratio of observed deaths to expected deaths. The likelihood ratio is a measure of how the mortality rate within a circle differs from the rate outside the circle and is given by the formula

$$\mathrm{LLR} = (O \ln(O/E)) + \{(C - O)\ln[(C - O)/(C - E)]\},$$

where LLR represents the logarithm of the likelihood ratio, O is observed deaths, E is expected deaths, and C is the total number of deaths in the entire analysis (522,994). This formula assumes that disease events are distributed as a Poisson random variable. While a non-logarithmic formulation is more commonly given in the literature (Kulldorff et al., 1997), for large data sets it requires storing intermediate values greater than 2^{1024}, which is the computational limit of most personal computers.

Likelihood ratios were compared to the results of a Monte Carlo simulation of the data, and circles with a likelihood ratio exceeding 95% of those obtained from the simulation were considered statistically significant and retained. The resultant circles, representing areas of both excess and deficit, were stratified into 10 levels of relative risk. Within each risk level, the circle with the highest likelihood ratio (lowest p value) was mapped. Circles with lower likelihood ratios were also mapped if they did not overlap any previously mapped circle within the same relative risk level. The entire analysis was performed using commercially available GIS software along with the Monte Carlo results from SaTScan.

Results

The method identified statistically significant excesses in mortality in a number of regions of the country, including much of the West, Upper Midwest, and Northeast (Fig. 1). Also identified were numerous subareas of higher relative risk that are significant under their own power, including three areas with a relative risk above 1.4, in Vermont, Iowa, and Wyoming-Idaho. In all, a total of 37 distinct elevated areas were mapped, nine of which would have been identified by SaTScan using the recommended maximum allowable circle size of 50% of the population at risk. The method also identified 35 distinct areas of mortality that were statistically significantly lower than expected, encompassing most of the South and South-Central states, New York City, and the extreme Southwest, four of which would have been identified by SaTScan using the same recommended maximum allowable circle size.

Included in the elevated areas is the Atlanta metropolitan region, where an excess of more than 10% in prostate cancer deaths was seen despite being part of a very large area of deficit that encompasses nearly the entire South. In addition, several clusters were identified specifically through the use of the grid-based approach, including a five county area of excess in southeastern Kansas and northeastern Oklahoma that is also located within a broader deficit region.

Discussion

Our approach provides more visual information than the results from SaTScan, while avoiding the visual chaos resulting from attempting to display millions of mostly similar circles. Comparable results could have been obtained through consideration of variable maximum allowable circle sizes or isotonic regression functions, but relative risk is a more familiar concept than either of these two approaches.

Comparing our statistical summary map with a conventional choropleth map of the same data as published in the *Atlas of Cancer Mortality* (Devesa et al., 1999) yields a number of compelling similarities as well as differences (Fig. 2). The statistical excesses found in North Dakota, Montana, Wyoming and surrounding states are quite evident in the choropleth map, as red ink far outweighs blue in this part of the country. While the population in most of these counties is very small, including a number that are classified as having sparse data, the sheer number of elevated counties in this region suggests a strong pattern. The same is true for the excesses seen in northern California and northern Vermont and New Hampshire. In contrast, the statistical excess we identified in the Maryland-Delaware area is much less apparent on the choropleth map. Conversely, the choropleth map hints at an apparent localized elevation in the Florida panhandle and south Georgia that is not borne out statistically. In terms of overall appearance, the statistical summary map resembles maps produced through conventional smoothing techniques (Kafadar, 1996; Kafadar, 1997). The extent to which a statistical summary map may confer advantages over a map generated through an empirical smoothing algorithm is a subject for future research.

One drawback of our method is that in attempting to correct for the tendency to display areas with large populations and large absolute excesses but small

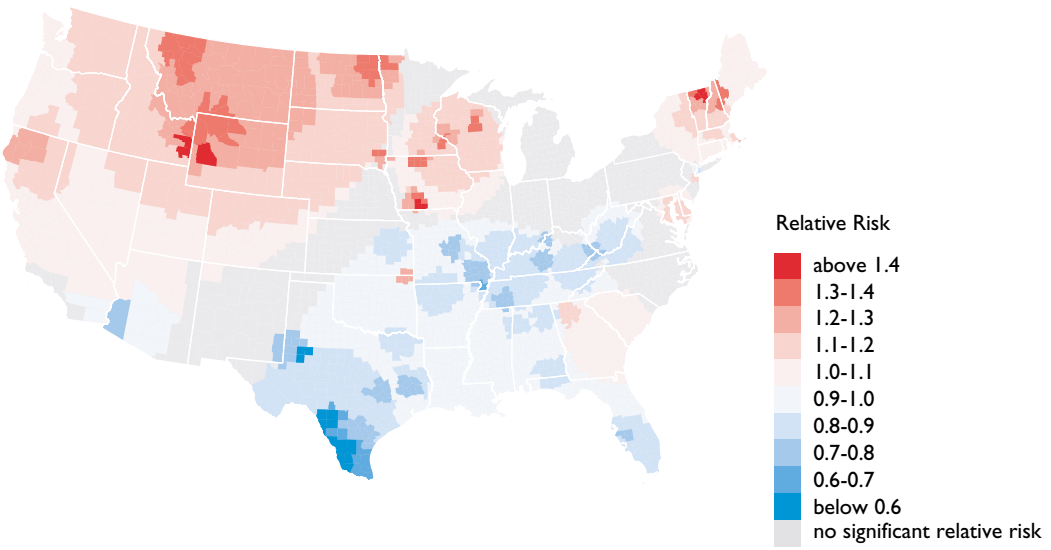

Fig. 1. Counties with statistically significant relative risks of prostate cancer mortality, white males, 1970–1994.

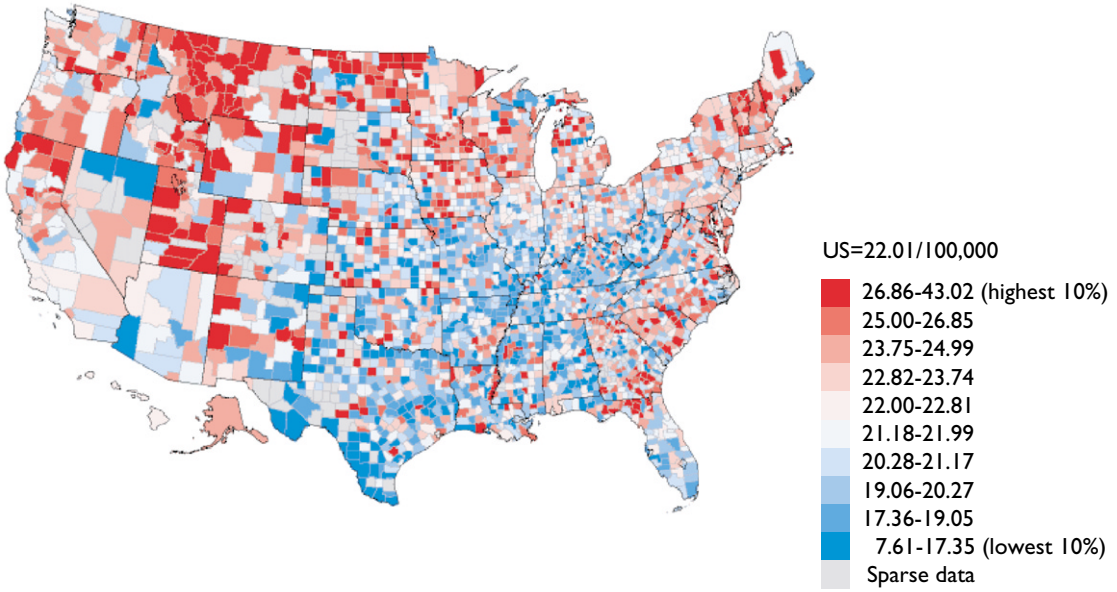

Fig. 2. Prostate cancer mortality rates by county, white males, 1970–1994.

elevated risks, a potential visual bias is produced in the opposite direction (Gelman and Price, 1999). Specifically, all the counties in the clusters with relative risks of 1.3 or higher are sparsely populated, and these are the counties which are most visually evident. It is far more likely to have a chance observation of 390 cases in a rural county where 300 are expected (relative risk = 1.30, log-likelihood ratio = 12.3, $p = 0.02$) than to observe 13,000 deaths in an urban county where 10,000 are expected (RR = 1.30, LLR = 419.5, $p < 0.0001$).

We find this to be a problem of aggregation bias (that is, the grouping of people into units of vastly differing sizes) rather than representing a deficiency in the method. For example, if the United States were to be divided into 10,000 districts such that each district would have expected approximately 50 prostate cancer deaths among white males between 1970–1994, then urban neighborhoods and rural counties would be equally likely to be highlighted with a given level of shading. Such an analysis is not practical with this data set, but many counties have census enumeration districts that are relatively homogeneous with respect to size and population characteristics. Analyses using data sets aggregated to such a level of geography would see this problem minimized.

The variation in prostate cancer mortality may provide clues to the etiology of prostate cancer, which is largely unknown. Hsing and Devesa (2001) hypothesized that agricultural exposures may contribute to the higher rates among White men in the north-central and western United States and among Black men in the southeastern areas of the country. Elevated risks for prostate cancer among farmers have been identified in numerous studies in the United States and Canada (Saftlas et al., 1987; Fincham et al., 1992; Cerhan et al., 1998; Parker et al., 1999; Buxton et al., 1999). Differences in screening, treatment and survival patterns, differential access to health care, or bias related to the attribution of prostate cancer as the underlying cause of death also provide alternative explanations for the geographic variation in prostate cancer mortality (Feuer et al., 1999; Lai et al., 2000, 2001a, b).

In conclusion, consideration of relative risk and statistical significance in combination conveys more information than either of these measures alone. Stratification by relative risk is a conceptually straightforward means of identifying sub-clusters. Displaying nested circles helps to convey the message that the mapped clusters do not have precise boundaries, but rather consist of a relatively well-defined core and less-certain periphery. This approach can also identify areas of significant localized excess contained within a broader region of deficit, as in the case of the Atlanta metropolitan region, or vice versa. Finally, considered at a sufficiently fine level of detail, this approach could be used as a means of visualizing a dose–response relationship from an environmental point source.

References

Besag, J., Newell, J., 1991. The detection of clusters in rare diseases. Journal of the Royal Statistical Society A 154 (1), 143–155.

Buxton, J.A., Gallagher, R.P., Le, N.D., Band, P.R., Bert, J.L., 1999. Occupational risk factors for prostate cancer mortality in British Columbia, Canada. American Journal of Industrial Medicine 35 (1), 82–86.

Cerhan, J.R., Cantor, K.P., Williamson, K., Lynch, C.F., Torner, J.C., Burmeister, L.F., 1998. Cancer mortality among Iowa farmers: recent results, time trends, and lifestyle factors (United States). Cancer Causes and Control 9 (3), 311–319.

Devesa, S.S., Grauman, D.J., Blot, W.J., Pennello, G.A., Hoover, R.N., Fraumeni, J.F., 1999. Atlas of cancer mortality in the United States, 1950–94. National Institutes of Health, Bethesda, MD. NIH Publication No. 99-4564.

Feuer, E.J., Merrill, R.M., Hankey, B.F., 1999. Cancer surveillance series: interpreting trends in prostate cancer—part II: cause of death misclassification and the recent rise and fall in prostate cancer mortality. Journal of the National Cancer Institute 91 (12), 1025–1032.

Fincham, S.M., Hanson, J., Berkel, J., 1992. Patterns and risks of cancer in farmers in Alberta. Cancer 69 (5), 1276–1285.

Forand, S.P., Talbot, T.O., Druschel, C., Cross, P.K., 2002. Data quality and the spatial analysis of disease rates: congenital malformations in New York State. Health and Place 8 (3), 191–199.

Fotheringham, A.S., Zhan, F.B., 1996. A comparison of three exploratory methods for cluster detection in spatial point patterns. Geographic Analysis 28 (3), 200–218.

Gelman, A., Price, P.N., 1999. All maps of parameter estimates are misleading. Statistics in Medicine 18 (23), 3221–3234.

Gregorio, D.I., Kulldorff, M., Barry, L., Samocuik, H., Zarfos, K., 2001. Geographical differences in primary therapy for early stage breast cancer. Annals of Surgical Oncology 8 (10), 844–849.

Hsing, A.W., Devesa, S.S., 2001. Trends and patterns of prostate cancer: what do they suggest? Epidemiologic Reviews 23 (1), 3–13.

Jemal, A., Devesa, S., Kulldorff, M., Hayes, R., Fraumeni, J., 2000. Geographic variation in prostate cancer mortality rates among white males in the United States. Annals of Epidemiology 10 (7), 470.

Kafadar, K., 1996. Smoothing geographical data, particularly rates of disease. Statistics in Medicine 15 (23), 2539–2560.

Kafadar, K., 1997. Geographic trends in prostate cancer mortality: an application of spatial smoothers and the need for adjustment. Annals of Epidemiology 7 (1), 35–45.

Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics: Theory and Methods 26 (6), 1481–1496.

Kulldorff, M., 1999a. Geographic information systems (GIS) and community health: some statistical issues. Journal of Public Health Management and Practice 5 (2), 100–106.

Kulldorff, M., 1999b. An isotonic spatial scan statistic for geographical disease surveillance. Journal of the National Institute of Public Health 48 (2), 94–101.

Kulldorff, M., Nagarwalla, N., 1995. Spatial disease clusters: detection and inference. Statistics in Medicine 14 (8), 799–810.

Kulldorff, M., Feuer, E.J., Miller, B.A., Freedman, L.S., 1997. Breast cancer clusters in the northeast United States: a geographic analysis. American Journal of Epidemiology 146 (2), 161–170.

Kulldorff, M., Rand, K., Gherman, G., Williams, G., DeFrancesco, D., 1998. Sat Scan v. 2.1: Software for the spatial and space-time scan statistic. National Cancer Institute, Bethesda, MD. On-line: http://dcp.nci.nih.gov/BB/Sat-Scan.html.

Kulldorff, M., Information Management Services, Inc., 2002. SaTScan v. 3.0: Software for the spatial and space–time scan statistics. National Cancer Institute, Bethesda, MD. On-line: http://srab.cancer.gov/satscan/download.html.

Lai, S., Lai, H., Krongrad, A., Lamm, S., Schwade, J., Roos, B.A., 2000. Radical prostatectomy: geographic and demographic variation. Urology 56 (1), 108–115.

Lai, S., Lai, H., Krongrad, A., Roos, B.A., 2001a. Overall and disease-specific survival after radical prostatectomy: geographic uniformity. Urology 57 (3), 504–509.

Lai, S., Lai, H., Lamm, S., Obek, C., Krongrad, A., Roos, B., 2001b. Radiation therapy in non- surgically treated nonmetastatic prostate cancer: geographic and demographic variation. Urology 57 (3), 510–517.

Le, N., Marrett, L.D., Robson, D.L., Semenciw, R.M., Turner, D., Walter, S.D., 1995. Canadian Cancer Incidence Atlas.. Ministry of Supply and Services, Ottawa, ON.

MacEachren, A.M., Brewer, C.A., Pickle, L.W., 1998. Visualizing georeferenced data: representing reliability of health statistics. Environment and Planning A 30 (9), 1547–1561.

National Cancer Institute, 2001. Cancer mortality maps & graph web site. On-line: http://www3.cancer.gov/atlasplus/.

Openshaw, S., Charlton, M., Wymer, C., Craft, A., 1987. A mark 1 geographical analysis machine for the automated analysis of point data sets. International Journal of Geographic Information Systems 1 (4), 335–358.

Parker, A.S., Cerhan, J.R., Putnam, S.D., Cantor, K.P., Lynch, C.F., 1999. A cohort study of farming and risk of prostate cancer in Iowa. Epidemiology 10 (4), 452–455.

Pickle, L.W., Mungiole, M., Jones, G.K., White, A.A., 1996. Atlas of United States mortality. National Center for Health Statistics, Hyattsville, MD.

Saftlas, A.F., Blair, A., Cantor, K.P., Hanrahan, L., Anderson, H.A., 1987. Cancer and other causes of death among Wisconsin farmers. American Journal of Industrial Medicine 11 (2), 119–129.

Sankoh, O.A., Yé, Y., Sauerborn, R., Müller, O., Becher, H., 2001. Clustering of childhood mortality in rural Burkina Faso. International Journal of Epidemiology 30 (3), 485–492.

Timander, L.M., McLafferty, S., 1998. Breast cancer in West Islip, NY: a spatial clustering analysis with covariates. Social Science and Medicine 46 (12), 1623–1635.

Turnbull, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L., Clark, L.C., 1990. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. American Journal of Epidemiology 132 (1), S136–143.

Walter, S.D., Birnie, S.E., 1991. Mapping mortality patterns: an international comparison. International Journal of Epidemiology 20 (3), 678–689.