TJC Turkish Journal of Cancer
Volume 34, No.2, 2004

Analyzing factors associated with cancer occurrence: A geographical systems approach

MUTLU HAYRAN

Hacettepe University Institute of Oncology, Department of Preventive Oncology, Ankara-Turkey

ABSTRACT

Geographical Information Systems (GIS) is defined as an automated set of functions that provides professionals with advanced capabilities for the storage, retrieval, display and analysis of geographically located data. This paper aims to show a GIS based approach to analysis of registry-based cancer incidence data. Case data were obtained from the Pennsylvania Cancer Registry for 1993-1997 period and corresponding population data were obtained from the National Center for Health Statistics. The incidence rate was calculated as the number of cases diagnosed per 100,000 persons living in any given area. The rates were age-adjusted (direct method) against the 1970 U.S. standard population. GIS methods such as clustering, autocorrelation were used to analyze prostate cancer incidence spatially. A lower incidence area in the northeast was observed. The significant variables that are associated with the incidence in this period were percentage of households with less than 2 bedrooms (p=0.022) and distance to toxic sites (p=0.007). These two factors were thought to represent different dimensions of urbanicity. GIS analyses for ecological analyses of health related outcomes is feasible and useful in generating hypothesis and identifying areas for intervention. [Turk J Cancer 2004;34(2):67-70]

KEY WORDS:

Geographical information systems, cancer incidence, prostate cancer

INTRODUCTION

Geographical Information Systems (GIS) is defined as an automated set of functions that provides professionals with advanced capabilities for the storage, retrieval, manipulation and display of geographically located data. With the development of GIS and the software to carry out its functions, federal, state and local health agencies are increasingly adapting their surveillance systems to capture address-based data for the location of health related events (1). This in turn provides opportunities to map the high and low rates of adverse health outcomes and detect areas where the health care the population receives may have an effect on these outcomes.

Although the temporal trends of cancer-related outcomes have been described and commented upon in detail, the geographical distribution of the disease has not been studied as much (2). A geographical systems approach may be useful in identifying factors affecting cancer morbidity. This paper analyzes the distribution of the prostate cancer-

related heath outcomes in space and tries to associate outcomes with the geographical distribution of several possible risk factors to help achieve this objective.

INTRODUCTION

Source of data

The source of the incidence data for prostate cancer is the Pennsylvania Cancer Registry. These data include information on prostate cancer cases diagnosed during the study period (1993-1997), information including age, race, residence, diagnosis date, stage at diagnosis and vital status. The geographic level of detail is the postal zip code for all cases. These data were used to calculate age-adjusted and age-specific prostate cancer incidence by race, stage, geographic location and calendar year. This data set was also used to calculate observed and relative survivals by year, age, race and cancer stage.

Population census counts

Population census counts include the estimated population counts for the period 1993-1997 stratified by age, sex, race, calendar year and geographical location obtained from the National Center for Health Statistics. These constitute the denominator data for incidence calculations.

Calculation of the incidence estimates

The incidence rate is the number of cases diagnosed per 100,000 persons living in any given area. The rates were age-adjusted (direct method) against the 1970 U.S. standard population. Case data were obtained from the Pennsylvania Cancer Registry for 1993-1997 period and corresponding population data were obtained from the National Center for Health Statistics.

Distance-based measures

Two distance based measures were used. These measures used two databases to calculate the Euclidean distances to the nearest hospital and the toxic release sites for each point on the Pennsylvania map. The first database, provided by the Pennsylvania Department of Health, includes the names and the geocoded locations of the state approved hospitals. The second one, the Toxic Chemical Release Inventory database originated by U. S. Environmental Protection Agency, shows the point locations of the sites releasing toxic substances in the state of Pennsylvania. Both databases were obtained from Pennsylvania Spatial Data Access (PASDA) library of the Pennsylvania State University.

GIS mapping

The first step in analyzing any given health outcome is to visualize its distribution over space. This is done by the mapping methods of GIS. There are various ways to present data on the maps. These include proportional circles, choropleth maps (presenting areas with different colors), contour maps, density equalizing maps and 3-D surface plots. Each of these techniques allows the user to assign different colors, shapes or sizes to the objects on the map or divide the area into zones depending on the value of the parameter it represents.

Exploratory analyses of spatial data sets are used to detect patterns. However, these analyses are sensitive to variations in the techniques used to determine the boundaries, the scale at which data are recorded, the sampling systems used to collect data, and the extent of the total area that is considered in the analysis (3). Smoothed maps are used to establish the stability of the estimated rates in each location by averaging them out using the rates observed in the neighboring locations.

Clustering

Clustering occurs if a disease event is seen more often in a particular area than would be expected by chance. The clusters of disease cases are meaningful only after having adjusted for spatial variations in the density of the background population. The methods exploring patterns of geographical clustering in disease may either look at overall clustering or they may try to detect location of clusters. The spatial scan statistics introduced by Kulldorff uses likelihood ratio testing and Monte Carlo simulations and was the method used here to identify the significantly low and significantly high clusters in the state of Pennsylvania (4).

Spatial autocorrelation

The word autocorrelation refers to the correlation of a variable across geographical location. The prefix ‘auto’ is

Table 1
Ordinary least squares estimation, factors affecting age-adjusted prostate cancer incidence, all races, Pennsylvania, 1993-97

Varitable	β	S.D.	t-value	p-value
Constant	134.85	12.82	10.52	<0.001
% households with <2 bedrooms	2.08	0.88	2.35	0.022
Distance to toxic sites	-92.24	33.26	-2.77	0.007

Akaike Information Criterion: 625.413. R^2: 0.1833, LM-lag: 0.18, LM-error: 0.36, Heteroskedasticity: 0.25

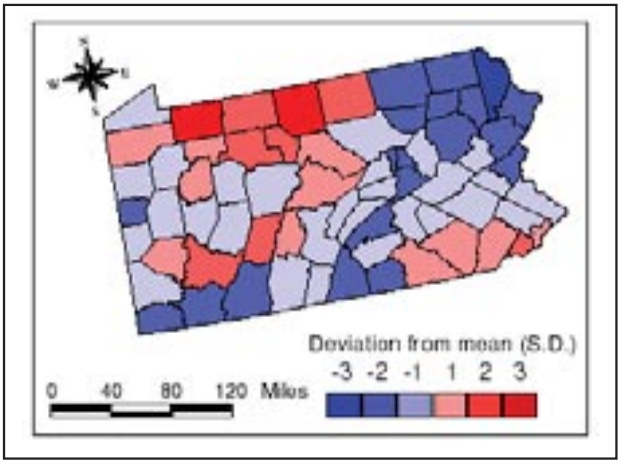

Fig 1. Smoothed maps for age-adjusted prostate cancer incidence, all race, Pennsylvania, 1993-97

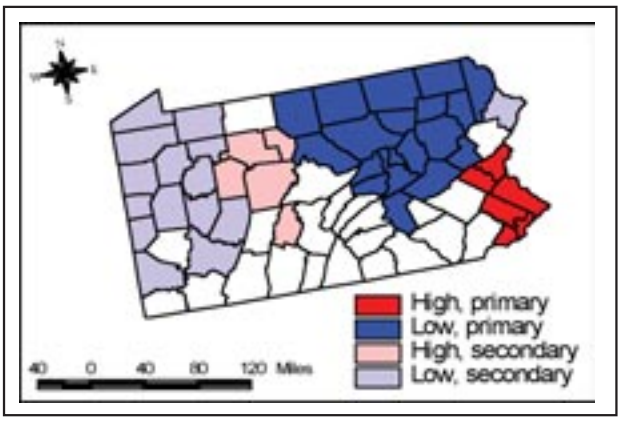

Fig 2. High and low clusters detected using spatial scan statistics, age-adjusted prostate cancer incidence, all race, Pennsylvania, 1993-97

used to indicate that this is a correlation of one variable with itself but not a correlation between two variables. The interdependence of variables can be used to elucidate disease patterns or to account for the effects of non-independence of closely positioned data points when exploring relationships between variables. Moran's I test is used here to investigate autocorrelations (5).

Spatial regression analyses

The regression analyses will show the results of the ordinary least squares (OLS) regression, along with the indicators for the fit (Akaike Information Criterion and R2) and the diagnostics for the presence of residual spatial autocorrelations (p-values for Lagrange Multiplier (LM) test for the spatial-lag and the spatial-error), which can affect the precision and the biasedness of the estimates (6,7). If these diagnostics indicate a need for the modification of the model, the results of the revised model will be presented.

RESULTS

The age-adjusted incidence rates ranged from 55.4 to 193.92 (IQR:24.1) per 100,000 with a median of 134 per 100,000 and the distribution was slightly skewed to the left. The global autocorrelation was not significant for the Moran's I statistic (I=0.084, p=0.17). A lower incidence area in the northeast was observed in the smoothed maps for the 1993-97 period (Figure 1). The same finding was observed as a primary low cluster to the northeast on the map showing the spatial scan statistic results (Figure 2). The primary high cluster on this map consisted of the Philadelphia, Delaware, Montgomery, Bucks, Lehigh and Northampton counties.

The significant variables that are associated with the incidence in this period were percentage of households with less than 2 bedrooms (p=0.022) and distance to toxic

sites (p=0.007). There were no residual autocorrelation in the OLS model nor any indication of heteroskedasticity, so no further models were studied.

DISCUSSION

A lower incidence area was observed to the northeast using both the smoothed maps and the Kuldorff technique. The Kuldorff method was also able to detect clusters when used to measure the level of clustering of leukemia in northern New York, to detect breast cancer clusters in northeast United States, and to analyze childhood leukemia data in Sweden (4,8,9).

The factors associated with incidence were percentage of households with fewer than 2 bedrooms and distance to toxic sites. Percentage of households with fewer than 2 bedrooms were positively associated with higher incidence, while the distance to toxic sites was negatively associated. The distance to toxic sites decreases by increasing urbanicity (r=-0.77) and, in these analyses, is thought to represent a measure of urbanicity. The increased incidence in urban areas can be explained by increased detection where people have easier access to comprehensive hospitals and have a higher chance of getting PSA screened at the time of an admission to the hospital with another genitourinary tract problems. This explanation is in part supported by the South Australian study, where Weller et al interviewed 3016 men in 1995 and showed that only experiencing lower urinary-tract symptoms influenced men to have PSA screening (10). In the same study, PSA screening rate was not associated with factors such as occupation or education.

CONCLUSION

Using geographical information systems analyses for ecological analyses of health related outcomes is feasible and useful in generating hypothesis and identifying areas for intervention. Regression analyses involving variables observed in geographical scales are prone to the violation of independent observations assumption due to the corraletedness of neighboring geographical locations, and such correlation structures can be sought and accounted for using existing spatial autoregressive models.

References

1. Fiore BJ, Hanrahan LP, Anderson HA. Public health response to reports of clusters. Am J Epidemiol 1990;132(Supp 1):S14-S21.
2. Ries LAG, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L, Edwards BK, editors. SEER Cancer Statistics Review, 1973-1997. National Cancer Institute, Bethesda, MD, 2000;393-404.
3. Fotheringam AS, Rogerson PA. GIS and spatial analytical problems. Int J Geog Inf Sys 1993;7:3-19.
4. Kulldorff M, Feuer EJ, Miller BA, et al. Breast cancer clusters in the northeast united states: a geographic analysis. Am J Epidemiol 1997;146:161-70.
5. Burridge P. On the Cliff-Ord test for spatial autocorrelation. J Roy Statist Soc 1980;42:107-8.
6. Griffith D. Practical Handbook of Spatial Statistics. New York: CRC Press, 1996.
7. Anselin L. Spacestat Tutorial: A workbook for using SpaceStat in the analysis of spatial data. Technical Software Series S-92-1, NCGIA, University of California, Santa Barbara, 1992.
8. Kulldorff M, Nagarwalla N. Spatial disease clusters: Detection and inference. Stat Med 1995;14:799-810.
9. Hjalmars U, Kulldorff M, Gustafsson G, et al. Childhood leukemia in Sweden; using GIS and a spatial scan statistic for cluster detection. Stat Med 1996;15:707-15.
10. Weller D, Pinnock C, Sialgh C, et al. Prostate cancer testing in SA men: influence of sociodemographic factors, health beliefs and LUTS. Aust N Z J Public Health 1998; 22S:400-2.