

Spatial analysis of childhood leukemia in a case/control study

Steve Selvin, Kathleen E. Ragland, Ellen Yu-Lin Chien, Patricia A. Buffler

University of California, Berkeley, USA

Received November 5, 2003 · Revision received May 11, 2004 · Accepted May 20, 2004

Abstract

A simple and direct analysis of the spatial distribution of childhood leukemia was performed using geographic data from a large case/control study. The data consist of cases of childhood leukemia and their corresponding birth cohort controls located in seven San Francisco Bay Area counties. Both parametric and randomization analyses show no evidence of a non-random spatial pattern of childhood leukemia among six of these counties. The data from San Francisco County, however, produce a moderately small significance probability (0.08) arising from a distance analysis and a significant p -value (0.01) arising from a frequency analysis of concordant case pairs. Although these p -values accurately reflect the probability of the observed spatial pattern occurring by chance alone, these results are based on only four cases of leukemia.

Key words: Case/control design – spatial analysis – childhood leukemia

Introduction

Geographic or spatial patterns of childhood leukemia have been occasionally reported (Cartwright et al., 2001; Dickinson et al., 2002; Land et al., 1984; Chen et al., 1997; Knox, 1994; Gustafsson and Carstensen, 2000; Pobel and Viel, 1997; Reynolds et al., 1996, 2002; Michelozzi et al., 2002; Grosche et al., 1999; Aickin et al., 1992; Alexander, 1992) but their statistical interpretation has not produced a consensus on the simple question: is there evidence of a non-random spatial pattern associated with the incidence of childhood leukemia? This is a question that was first posed in the 1950's and has yet to be resolved. It is recognized that acquiring some understanding of the temporal and spatial patterns of leukemia may provide important etiologic insights. Thus, recent studies in the United Kingdom have attempted to address this

need. The data and analyses that follow address this same question using recently collected cases of childhood leukemia and controls in the United States.

Methods

Data

Case/control data collected from the San Francisco Bay Area (1995-1999) provide an opportunity to explore the spatial distribution of childhood leukemia using relatively large numbers of observations (Table 1).

These case/control data originate from the Northern California Childhood Leukemia Study (NCCLS) and are a small part of this large and far-ranging study of childhood leukemia (Ma, X. et al., 2002). The specific observations consist of the addresses of 112 leukemia cases and 221

Corresponding author: Steve Selvin, School of Public Health, University of California, Berkeley, Berkeley, CA 94720, USA. e-mail: selvin@stat.berkeley.edu, phone: 001 510 642 4618

Table 1. Populations, areas, number of cases and number of controls from seven San Francisco Bay area counties.

county	populations*	areas**	cases	controls	totals
Alameda	310,916	1952.7	40	77	117
Contra Costa	219,550	1963.5	20	40	60
Marin	43,742	1407.5	5	7	12
San Francisco	97,852	122.3	4	9	13
San Mateo	141,032	1193.9	11	19	30
Santa Clara	362,874	3376.2	22	51	73
Sonoma	96,034	4114.2	10	18	28
combined	1,272,000	14,131.2	112	221	333

* = children ages 0–15 from 2000 US census counts.
 ** = area of county in square kilometers.

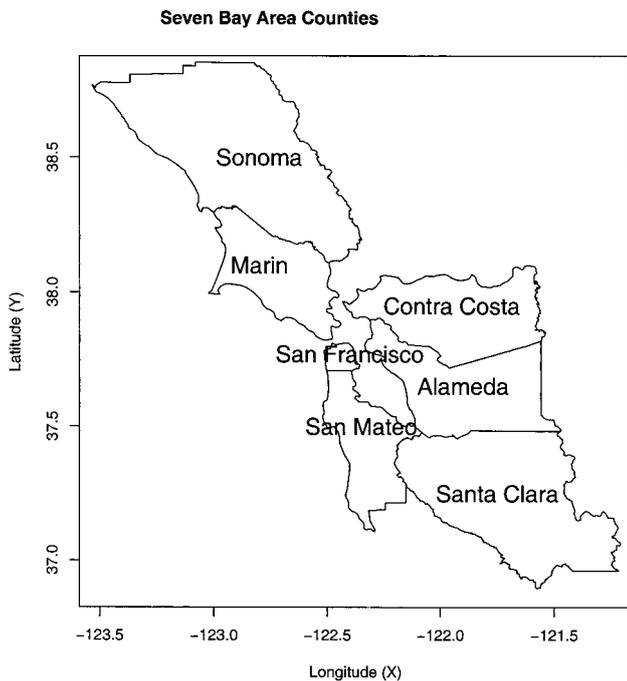


Fig. 1. The seven San Francisco Bay Area counties that make up the geographic study area.

birth controls from seven contiguous San Francisco Bay Area counties (Alameda, Contra Costa, Marin, San Francisco, San Mateo, Santa Clara and Sonoma – Table 1 and Figure 1). The cases are children (ages 0 to 14) with newly diagnosed leukemia during the years 1995 to 1999 ascertained from four major clinical centers in the greater Bay Area. Although the case ascertainment is hospital-based, a comparison with cases ascertained by the California State Cancer Registry shows that the NCCLS identified more than 90% of eligible children in the San Francisco metropolitan area (the five counties in the registry area and data was not available from two non-participating counties). Controls for the NCCLS were matched to these cases by date of birth. Specifically, the controls were identified by randomly selecting four to eight birth certificates of children born on the same day

that matched on sex, race, county of birth and Hispanic status of mother and father. Four to eight controls per case were selected because it was anticipated that some families would not agree to participate in the broader NCCLS. Essentially all cases and controls selected, however, were located geographically using birth address and whether or not they agreed to participate further in the study, and used in this spatial analysis. The residence of the mother at the time of the child’s birth was geographically located using a standard global positioning system (GPS) based on visiting the address of the study subject; thus producing a latitude and longitude measurement for all cases and controls. The final data set consists of 333 sets of matched pairs.

Figure 2 displays the spatial distribution of cases (dots) and controls (circles) for a single county (Alameda) to illustrate the data available to explore possible differences in spatial patterns. In addition, the latitude and longitude coordinates were mathematically transformed so that distance is measured in kilometers; that is, a new Cartesian coordinate system was established in kilometers relative to the latitude and longitude point (37.5, –122.5). Nearest neighbor distances were then calculated to compare statistically the spatial patterns of cases and controls.

Analysis

The nearest neighbor distance for a specific observation is simply the distance to the closest observation among the other case and control study participants. Nearest neighbor distances were classified into two categories – distances between two cases (case/case pairs) and distances between case/control or control/control pairs (jointly referred to as non-case/case pairs).

The nearest neighbor distance between case/case nearest neighbor pairs and the frequency of case/case pairs provide two statistical measures of non-randomness of the spatial data. The distribution of case/control and control/control pairs of nearest neighbor distances reflects the spatial pattern unaffected by case status. Thus, when leukemia influences the spatial location of the cases, a reduced mean nearest neighbor distance between case/case pairs should be observed as well as an elevated frequency of nearest neighbor case/case pairs relative to the non-case/case pairs.

When no spatial pattern exists, the mean nearest neighbor distances calculated from case/case and non-

Case/Controls in the Alameda County

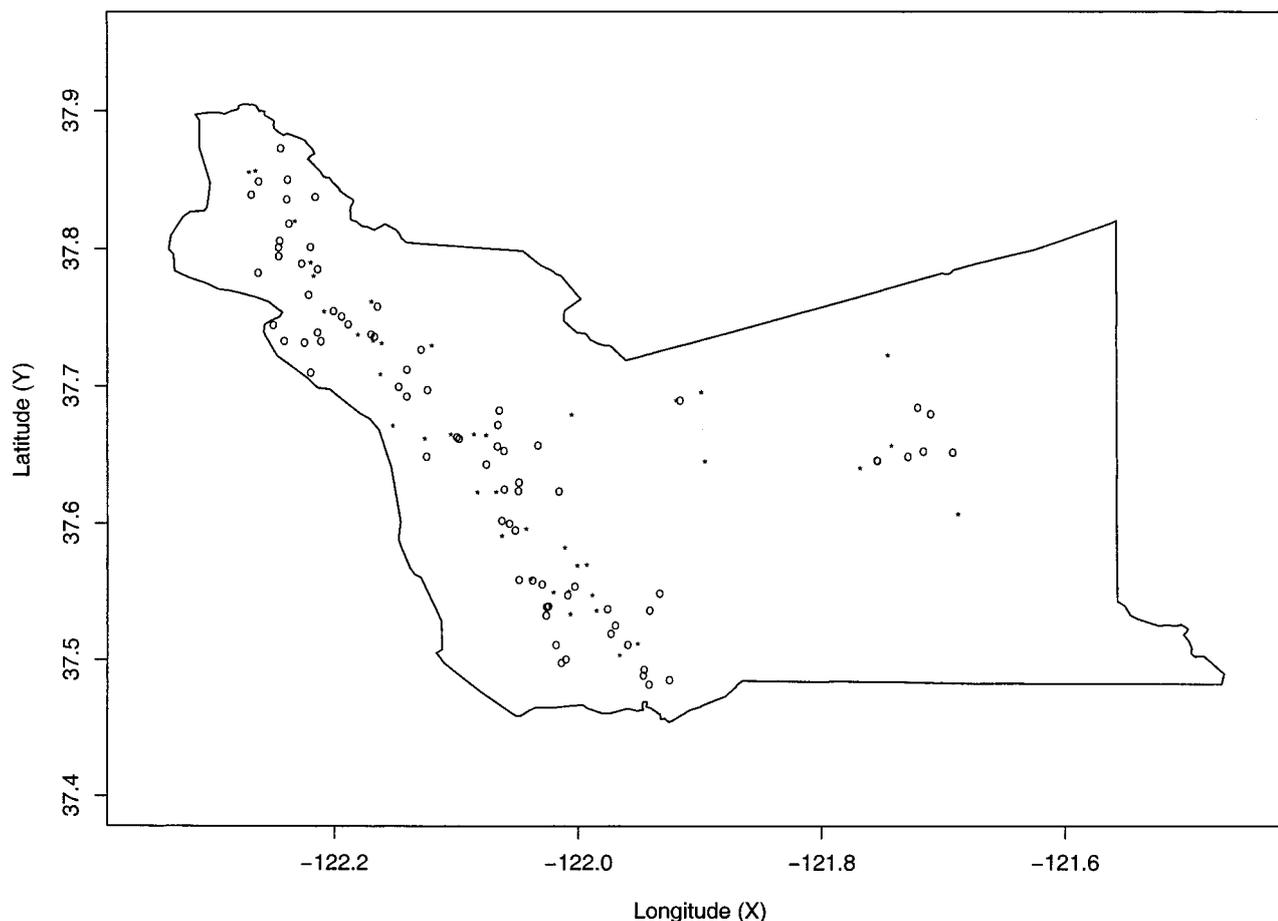


Fig. 2. The spatial distribution of the 40 cases of childhood leukemia and 77 controls in the county of Alameda, CA (1995–1999).

case/case pairs are expected to be equal and the frequency of the case/case pairs is expected to be equal to a known value that depends only on the number of cases and controls sampled.

The two kinds of nearest neighbor mean distances are compared using a typical two-sample test-statistic

$$z = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{\text{variance}(\bar{x}_1 - \bar{x}_0)}}$$

where \bar{x}_1 and \bar{x}_0 represent the mean nearest neighbor distances between case/case and non-case/case pairs, respectively.

This test-statistic z has an approximate standard normal distribution when no leukemia related systematic differences exist among mean nearest neighbor distances.

The observed frequency of case/case pairs of nearest neighbors (denoted m) is compared to the expected frequency calculated under the hypothesis that no leukemia related spatial pattern exists.

This approach is the topic of an extensive theoretical paper that explores the properties and power of using the number of case/case pairs as a test-statistic (Cuzick and Edwards, 1990).

That is, the value m is compared to its expected value (denoted $E[m]$)

$$\text{expected value} = E[m] = np = n \frac{\binom{n_1}{2}}{\binom{n}{2}} = \frac{n_1(n_1 - 1)}{n(n - 1)}$$

where p represents the probability of the occurrence of a case/case pair, n_1 represents the total number of cases and n represents the number of collected observations (controls + cases = $n_0 + n_1 = n$). The test-statistic used is

$$z = \frac{m + 0.5 - E[m]}{\sqrt{np(1 - p)}}$$

and it also has an approximate standard normal distribution when no leukemia related spatial pattern exists.

Case/Controls in the Seven Bay Area Counties

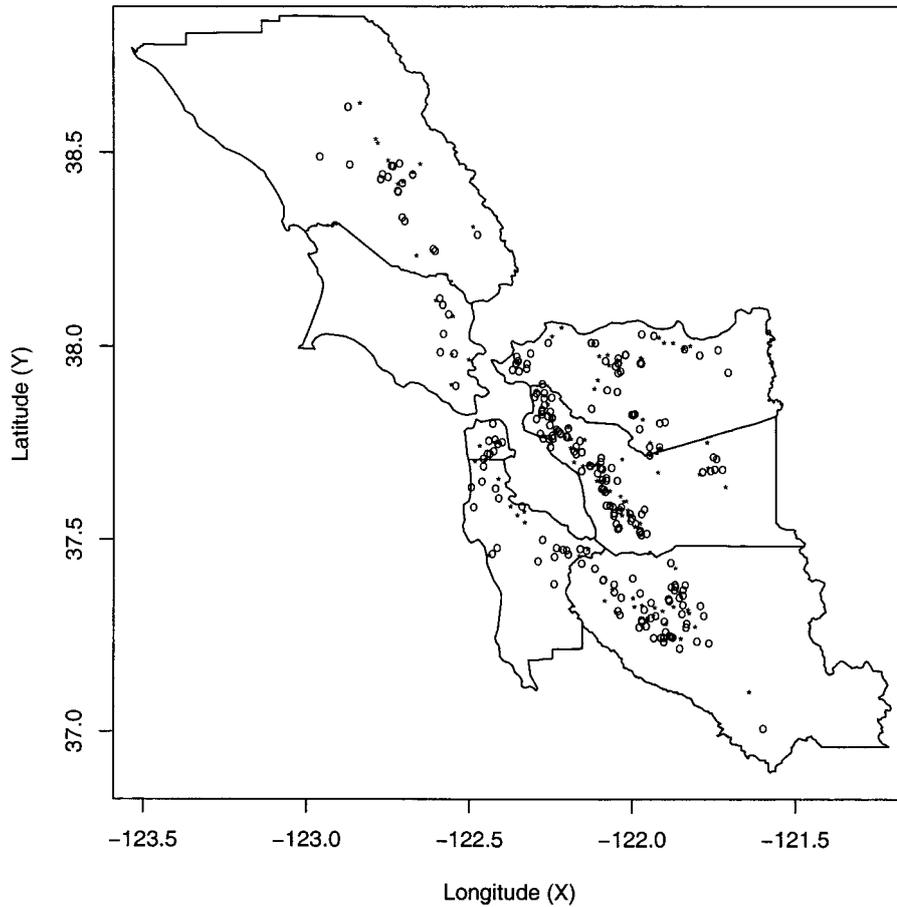


Fig. 3. The spatial distribution of the 112 cases of childhood leukemia and 221 controls in the seven Bay Area counties (1995–1999).

These two parametric analyses are supplemented by parallel randomization procedures applied to the two nearest neighbor measures that are entirely assumption-free and accommodate nearest neighbor statistics calculated for any number of observations and any shaped region (Besag and Diggle (1977). Significance levels (*p*-values) are estimated from the computer generated randomization distributions to evaluate the observed differences in mean nearest neighbor distances and the frequency of case/case pairs. That is, case and control status is assigned at random, producing randomized “data” that is used to estimate the distribution of these two test-statistics when only random differences exist between case and control spatial patterns. Significance probabilities (*p*-values) are then calculated from these estimated “null” distributions.

Results

Using the 333 case/control observations, two mean values are calculated: the mean nearest neighbor distance between case/case pairs (\bar{x}_1) and the mean nearest neighbor distance between the non-case/case pairs (\bar{x}_0). These mean values along with their standard errors are presented in Table 2 for the six Bay Area counties and all the counties combined. Figure 3 displays the geographic locations of these 333 cases and controls for all counties.

For five counties and all counties combined the comparisons show no evidence of a non-random spatial pattern of childhood leukemia (no extreme *p*-values – Table 2). The County of San Francisco shows some evidence of a non-random spatial distribution but involves only three case/case pairs among the 13 observations.

Table 2. Nearest neighbor mean distances (in kilometers) for case/case and non-case/case pairs.

	case/case	\bar{x}_1 (s.e.)	non-case/case	\bar{x}_0 (s.e.)	p -values*
Alameda	12	1.360 (0.388)	105	1.027 (0.083)	0.89 (0.88)
Contra Costa	7	2.055 (0.444)	53	1.508 (0.201)	0.83 (0.79)
San Francisco	3	0.410 (0.038)	10	1.597 (0.411)	0.08 (0.11)
San Mateo	3	2.563 (0.234)	27	2.390 (0.332)	0.57 (0.52)
Santa Clara	7	1.046 (0.305)	66	1.760 (0.247)	0.18 (0.15)
Sonoma	3	1.880 (0.587)	25	2.035 (0.437)	0.45 (0.33)
combined	34	1.454 (0.198)	299	1.529 (0.091)	0.39 (0.37)

*one sided p -values for parametric and randomized test (in parentheses) procedures

Note: Marin county has no case/case nearest neighbor pairs and in San Francisco count three case/case pairs are so close in proximity that they are not distinguishable in Figure 3.

Table 3. Frequencies of case/case pairs, expected values and p -values.

	n	case/case (m)	expected ($E[m]$)	p -values*
Alameda	117	12	13.45	0.61 (0.61)
Contra Costa	60	7	6.44	0.33 (0.31)
Marin	12	0	1.82	0.86 (0.82)
San Francisco	13	3	1.00	0.01 (0.01)
San Mateo	30	3	3.79	0.56 (0.56)
Santa Clara	73	7	6.42	0.33 (0.33)
Sonoma	28	3	3.33	0.46 (0.46)
combined	333	34	37.45	0.70 (0.70)

*one sided p -values for parametric and randomized test (in parentheses) procedures

To supplement the parametric approach, a parallel randomization procedure is used to evaluate the observed differences between the seven mean nearest neighbor distances. Assigning case/control status at random to the observed spatial locations produced 1000 replicate “case/control” mean nearest neighbor distances for each county that differ only because of sampling variation. The proportion of these null hypothesis generated differences in nearest neighbor means that are less than the original nearest neighbor mean values ($\bar{x}_1 - \bar{x}_0$) is used to estimate the significance levels. This assumption-free randomization approach essentially duplicates the two-sample parametric analysis (p -values in parentheses in Table 2).

Two issues arise in the evaluating of these results: the assumption that the nearest neighbor distances have at least an approximate normal distribution and that these distances are independently sampled. Mean nearest neighbor distances based on more than six observations show no evidence of a misleading non-normality (Donnelly, 1978) due primarily to the fact that nearest neighbor distances have an approximately symmetric (mean \approx median) and normal-like distribution. However, the independence of the observations remains a concern.

It is likely that searches for nearest neighbors involve overlapping areas causing a degree of non-independence.

This theoretical lack of independence has been shown to have little impact on evaluating nearest neighbor mean distances (Diggle 1978). Also, the similarity of the parametric and randomization analyses indicates that any lack of normality or non-independence of the nearest neighbor distances has no discernible systematic influence on the analysis of these data.

In addition to comparing the mean nearest neighbor distances, the frequency of case/case pairs also provides an assessment of the non-randomness of spatial patterns. For example, the data from Alameda county where $m = 12$ case/case pairs were recorded and $np = 117(0.115) = 13.45$ cases/case pairs are expected, a statistical test produces a p -value of 0.61, again using a normal distribution based test-statistic. The results from the same analysis applied to the six other counties and all seven counties combined are given in Table 3.

These comparisons of expected and observed values also show no important indications of a non-random spatial pattern of childhood leukemia with again the possible exception of San Francisco County.

In addition to a normal distribution based approximate approach, a randomization procedure was used to generate 1000 random values of the “case/controls pairs” for each county when no difference

Table 4. Standard errors for evaluation of the frequency of case/case pairs (m versus $E[m]$).

	randomization	theoretical ¹⁵	independence*
Alameda	3.311	3.396	3.450
Contra Costa	2.394	2.390	2.398
Marin	1.438	1.221	1.242
San Francisco	0.938	0.876	1.006
San Mateo	1.695	1.791	1.820
Santa Clara	2.419	2.466	2.419
Sonoma	1.725	1.702	1.714
combined	5.685	5.673	5.759

* = standard errors based on assumed independence of observations and $variance = np(1-p)$ where p again represents the probability of the occurrence of a case/case pair.

exists in spatial patterns. The proportion of randomized “case/case” pairs that exceeded the original value of m was computed (p -value). These estimated p -values produce results that hardly differ from the parametric approach (p -values in parentheses – Table 3).

The estimated standard errors for the randomization procedure, the standard errors calculated from entirely theoretical considerations (Michelozzi et al., 2002) and the standard errors based on the null hypothesis generated expected values do not substantially differ (Table 4).

When the estimates of variability do not differ, the three analytic approaches for assessing the case/case pair frequencies will not substantially differ.

Discussion

Both parametric and randomization analyses show no evidence of a non-random spatial pattern of childhood leukemia cases in this approximate population-based series of 333 cases and controls.

That is, the combined analysis indicates no evidence of a spatial pattern. The same analyses, in addition, applied to each county (at reduced power) also show no indication of important non-random case/control differences. The data from San Francisco County produce a moderately small significance probability (0.08) arising from the distance analysis (Table 2) and a significantly small p -value (0.01) arising in the frequency analysis (Table 3). Although these p -values accurately reflect the probability of the observed spatial pattern occurring by chance alone, the results from San Francisco County are based on observing only four cases of leukemia (three case/case pairs). For such a small number of observations, classification errors or even slight biases in reporting will likely have an extreme

impact on the analysis of a spatial pattern. A small error in the location or a single misclassification of a case, for example, would undoubtedly produce a considerable change in the statistical results.

Most epidemiologic case/control studies of childhood leukemia face the problem of differing participation rates between case and control groups, producing a bias that potentially influences subsequent analytic results. A feature of the NCCLS spatial data is the absence of “participation bias” due to the inclusion of essentially all selected case/control subjects. Another potential source of bias that is not an issue is the choice of the study population. Frequently the motivation to study a population comes from an observed cluster of leukemia cases. For example, the studies in Seascale, England (Aickin et al., 1992), Fallon, Nevada (Besag and Diggle, 1977) and the California Central Valley (Diggle, 1978) all followed a reported “cluster” of leukemia cases. The NCCLS data were collected as part of a study focused on a wide range of questions concerning childhood leukemia in a large and diverse population with no history of any unusual spatial patterns.

Most geographic studies use location of the residence at diagnosis and hence are population-based for incidence. By using birth residence for the NCCLS data, cases born during the same years as the controls but diagnosed prior to 1995 are excluded by study criteria. However, only a few such cases were omitted and no trends in leukemia risk are sufficiently strong to produce a meaningful bias from these excluded cases.

It should be noted that the NCCLS data were collected in a matched design which was not taken into account in the spatial analyses.

The bias incurred is small and negligible because the matching variables (age, sex, county and race) are not strong confounding factors. The gain in power from an unmatched analysis over a matched analysis is considerable (333 cases/control observations versus 150 pairs). In addition, less powerful matched analyses (not given) produced results that were similar to those that ignored the matched structure of the data collection.

One-sample nearest neighbor analyses are frequently biased by “edge-effects” (Reynolds et al., 2002) which arises because the theoretical nearest neighbor mean distance and its associated variance are generated based on regions without regard to boundaries whereas spatial data are typically restricted to well defined geographic areas causing a usually slight bias.

Because both cases and controls in the present analysis are subject to the same bias, they will be

similarly affected by the absence of data beyond the county boundaries. Thus, case/control comparisons will be unbiased and correction is unnecessary.

In general, a statistical approach that reduces a two-dimensional distribution to a one-dimensional summary incurs a loss of "information". Consequently, certain spatial configurations are not easily detected with specific spatial summary statistics such as near neighbor distances (low statistical power). It is, however, likely that many such patterns would be noticed by inspection, which then serves as a guide to the selection of a more powerful statistical measure. As always, the degree of statistical power associated with a summary of a geographic distribution of disease data depends largely on the postulated spatial pattern underlying the observed locations.

The controls for the spatial analysis of the NCCLS data were randomly selected from birth certificates but, in general, a control group can be selected from available US Census Bureau data and maps (e.g., TIGER files). It is relatively straightforward to select randomly controls from a series of census tracts with probabilities equal to age-, race- and sex-specific population counts using readily available US census data. Alternatively, census based block groups could be used to provide an even better approximation to the control population distribution. It is equally straightforward to select a random location within each of these geographic areas producing a set of spatially random control "observations". The implicit assumption in this process is that non-diseased individuals have a stable and uniform distribution within each selected area. The fact that people live in small clusters (single residences, apartment houses, hotels, along roads, etc.) and are not at random geographic locations is ignored when a randomly selected point represents the location of a control. This degree of approximation, however, is likely sufficient to estimate accurately the spatial distribution of most US populations.

Another notable issue that emerges from the analysis of the NCCLS spatial data is the accuracy of simple normal distribution based test-statistics. Comparing nearest neighbor distances and counts of case/case pairs using elementary parametric statistical tools produces analytic results that differ little from the computer intensive nonparametric approaches. Because of the availability of "control observations" and elementary statistical methods to compare spatial patterns, data collected using a case/control design should be useful for investigating the spatial patterns of a wide variety of human diseases.

The study was supported by two research grants from the Environmental Health Sciences, United State (R01 ES09137 and PS42 ES04705).

References

- Aickin, M., Chapin, C. A., Flood, T. J., Englender, S. J. and Caldwell, G. C.: Assessment of the spatial occurrence of childhood leukemia mortality using standardized rate ratios with a simple linear Poisson model. *Int J Epidemiol* 21, 649–65 (1992).
- Alexander, E. F.: Space-time clustering of childhood acute lymphoblastic leukemia: indirect evidence for transmissible agent. *Br J of Cancer* 65, 589–592 (1992).
- Besag, J. and Diggle, P. J.: Simple Monte Carlo test for spatial pattern. *Appl Statist* 26(3), 327–33 (1977).
- Cartwright, R. A., Dovey, G. J., Kane, E. V. and Gilman, E. A.: The onset of the excess of childhood cancer in Seascale, Cumbria. *J Public Health Med* 23(4), 314–22 (2001).
- Chen, R., Iscovich, J. and Goldbourt, U.: Clustering of leukemia cases in a city in Israel. *Stat Med* 16(16), 1873–87 (1997).
- Cuzick, J. and Edwards, R.: Spatial clustering for inhomogeneous populations. *J R Statist Soc B* 52(1), 73–104 (1990).
- Dickinson, H. O. and Parker, L.: Leukemia and non-Hodgkins lymphoma in children of male Sellafield radiation workers. *Int J Cancer* 99(3), 437–44 (2002).
- Diggle, P. J.: Note on Clark and Evans test of spatial randomness. In *Simulation methods in archaeology*, ed I. Hopper. Cambridge: Cambridge University Press 246–248 (1978).
- Donnelly, K. P.: Simulations to determine the variance and edge effect of total nearest neighbor distance. In *Simulation methods in archaeology*, ed I. Hopper. Cambridge: Cambridge University Press 91–95 (1978).
- Grosche, B., Lackland, D., Mohr, L., Dunbar, J., Nicholas, J., Burkart, W. and Hoel, D.: Leukemia in the vicinity of two tritium-releasing nuclear facilities: a comparison of the Kruemmel Site, Germany, and the Savannah River Site, South Carolina, USA. *J Radiol Prot* 19(3), 243–252 (1999).
- Gustafsson, B. and Carstensen, J.: Space-time clustering of childhood lymphatic leukemia and non-Hodgkins lymphomas in Sweden. *Eur J Epidemiol* 16(12), 1111–1116 (2000).
- Knox, E. G.: Leukemia clusters in childhood: geographical analysis in Britain. *J Epidemiol Community Health* 48(4), 369–376 (1994).
- Land, C. E., McKay, F. W. and Machado, S. G.: Childhood leukemia and fallout from the Nevada nuclear tests. *Science* 223(4632), 139–144 (1984).
- Ma, X., Buffler, P. A., Selvin, S., Wiencke, J. L., Wiemels, J. L. and Reynolds, P.: Day-care attendance and

- the risk of childhood acute lymphoblastic leukemia. *Br J of Cancer* 86, 1419–1424 (2002).
- Michelozzi, P., Capon, A., Kirchmayer, U., Forastiere, F., Biggeri, A., Barca, A. and Perucci, C. A.: Adult and childhood leukemia near a high-power radio station in Rome, Italy. *Am J Epidemiol* 155(12), 1096–1103 (2002).
- Pobel, D. and Viel, J. F.: Case-control study of leukemia among young people near La Hague nuclear re-processing plant: the environmental hypothesis revisited. *BMJ* 314(7074), 101–106 (1997).
- Reynolds, P., Smith, D. F., Satariano, E., Nelson, D. O., Goldman, L. R. and Neutra, R. R.: The four county study of childhood cancer: clusters in context. *Stat Med* 15(7–9), 683–697 (1996).
- Reynolds, P., Von Behren, J., Gunier, R. B., Goldberg, D. E., Hertz, A. and Harnly, M. E.: Childhood cancer and agricultural pesticide use: an ecologic study in California. *Environ Health Perspectives* 110(3), 319–324 (2002).