Exploring Associations between Residential Location and Breast Cancer Incidence in a Case–Control Study

Christopher Paulu,[1] Ann Aschengrau,[2] and David Ozonoff[1]

[1]Department of Environmental Health and [2]Department of Epidemiology and Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA

Locating geographic hot spots of cancer may lead to new causal hypotheses and ultimately to new knowledge of cancer-causing factors. The Cape Cod region of Massachusetts has experienced elevated incidence of breast cancer compared with statewide averages. The origins of the excess remain largely unexplained, even after the Upper Cape Cod Cancer Incidence Study investigated numerous potential environmental exposures. Using case–control data from this study (258 cases and 686 controls), we developed an exploratory approach for measuring associations between residential location and breast cancer incidence, adjusting for individual-level risk factors. We measured crude and adjusted odds ratios over the study region using fixed-scale grids and a smoothing algorithm of overlapping circular units. Polycircular hot spot regions, derived from the peak values of the smoothed odds ratios, delineated geographic areas wherein residence was associated with 60% [odds ratio (OR), 1.6; 95% confidence interval (CI), 0.8–3.2] to 210% (OR, 3.1; 95% CI, 1.3–7.2) increased incidence relative to the remainder of the study population. The findings suggest several directions for further research, including the identification of potential environmental exposures that may be assessed in forthcoming case–control studies. *Key words*: breast cancer, case–control studies, cluster analysis, epidemiologic methods, spatial analysis. *Environ Health Perspect* 110:471–478 (2002). [Online 1 April 2002]
http://ehpnet1.niehs.nih.gov/docs/2002/110p471-478paulu/abstract.html

The Cape Cod region of Massachusetts has experienced elevated incidence of cancer compared with statewide averages. At the public's request, Massachusetts funded a population-based case–control study of a five-town area of Cape Cod. The Upper Cape Cod Cancer Incidence Study (*1*) examined cancer in relation to numerous hypothesized environmental factors, including perchloroethylene-contaminated drinking water (*2–5*), proximity to cranberry cultivation (*6*), and potential exposure to airborne combustion by-products due to the burning of military munitions waste (*7*). However, the investigated exposures afforded only a partial explanation of the excess cancer incidence. In response, we conducted an exploratory spatial analysis of breast cancer incidence and residential location. Our goal was to describe associations between residential location and breast cancer incidence at geographic scales finer than the study region as a whole. The rationale was that identifying localized hot spots may lead to the identification of causal hypotheses.

Although much research has been done to analyze spatially the occurrence of cancer and other health outcomes (*8–15*), methods taking full advantage of the characteristics of case–control data have not been applied. To investigate the relation between breast cancer incidence and residential location with data from the Upper Cape Cod Cancer Incidence Study, we developed an exploratory model designed to map relative incidence rates by residential location, using an individual level of analysis and taking account of individual risk factors.

Methods

Selection and enrollment of study population. Cases were all incident cancers of the breast (n = 334) diagnosed from 1983 through 1986 among permanent residents of five towns in the Upper Cape Cod area of Massachusetts (Figure 1) and reported to the Massachusetts Cancer Registry. We selected controls from among demographically similar permanent residents of the Upper Cape Cod towns during the years from 1983 to 1986. We needed three sources to identify controls efficiently, because many cases were elderly or deceased when the study began. We chose living controls less than 65 years old using random-digit dialing, and those 65 years and older randomly from lists of Medicare beneficiaries furnished by the Health Care Financing Administration (HCFA). Deceased controls of ages similar to those of deceased cases we chose randomly from a file furnished by the Massachusetts Department of Vital Statistics and Research. The three sources of controls are considered to be complete with respect to their target populations (< 65 years old, ≥ 65 years, and deceased). Random selection within each of these sources should not be systematically biased with respect to residential location. The methods of subject selection are reported in more detail elsewhere (*3*).

The selected controls served as a source population of controls for nine cancer types in the Upper Cape Cod Cancer Incidence Study. Overall, we interviewed 79% of the cases, 76% of HCFA controls, 79% of deceased controls (we interviewed a proxy on behalf of subjects who were deceased), and 74% of contacted and eligible random-digit dial controls (Table 1). The demographic characteristics of interviewed and noninterviewed subjects were similar. We selected the control group for the breast cancer analysis by stratifying the breast cancer cases on the basis of age (in decades), vital status, and, if deceased, year of death, and then choosing all female controls who fell into a stratum with at least one case, yielding 763 controls (Table 2). We randomly assigned index years (comparable with year of diagnosis) to the controls in a weighted design to achieve identical distributions of diagnosis and index years. We excluded controls who moved to the Upper Cape Cod area after the index year (n = 46), as well as cases (n = 7) and controls (n = 31) with incomplete residency histories, leaving 258 breast cancer cases and 686 controls for the final analysis.

Mapping and digitizing of residency history. We asked subjects, or their proxies, to recall places and calendar years of residence in the five-town study area, dating back as far as 1943 (40 years before the earliest diagnosis year). We transcribed full addresses with street name, number, and village or township, or in some cases street names alone, cross streets, or nearest landmarks, and used tax assessors' books to determine the parcel of land corresponding to a street address. We then recorded residential location on a set of

Address correspondence to C. Paulu, Environmental Toxicology Program, Bureau of Health, State of Maine, Key Plaza, 8th Floor, 11 State House Station, Augusta, ME 04333-0011 USA. Telephone: (207) 287-9932. E-mail: cpaulu@mac.com

We gratefully acknowledge the study participants, who gave their valuable time and the benefit of their experience. We also thank P. Cyr and C. Barsotti for their assistance in digitizing location data, and N. Maxwell, R. Clapp, and D. Kriebel for their comments on the manuscript.

This publication was made possible by grant 2P42 ES07381 from the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), with funds from the U.S. Environmental Protection Agency (U.S. EPA). Support for this work was also provided by Silent Spring Institute with funds appropriated by the Massachusetts legislature and administered by the Massachusetts Department of Public Health (DPH). The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, NIH, U.S. EPA, Massachusetts DPH, or Silent Spring Institute.

Received 10 July 2000; accepted 16 November 2001.

enlarged paper reproductions of 1:24,000-scale U.S. Geological Survey maps. This enabled us to map complete addresses to within an estimated accuracy of 100 meters. We mapped all subjects without knowledge of case status.

The original database of residency information was on paper maps because the Upper Cape Cod Cancer Incidence Study was conducted before Geographic Information Systems were commonplace. For the present investigation, we transferred residential locations from the paper database to digital form using a large (36 × 48-in) digitizing board.

Exploratory model for studying spatial associations using individual-level data. To explore associations between residential location and cancer status, we defined independent "exposure" variables by dividing the study area into subregions. Each subregion ($s_1 \ldots s_n$) became a dichotomous (yes/no) "regional membership" attribute for individual subjects. For example, a "yes" value for s_1 meant that the subject had lived within the bounds of region s_1. Implementing this exploratory model required choosing a procedure for regionalizing the data (i.e., defining boundaries for $s_1 \ldots s_n$); a type of effect measure; a reference population; and a method for data visualization.

We developed two approaches, using different procedures for regionalizing the data. The first method used a set of regularly shaped grids (multiscale grids), whereas the second method used overlapping circles (adaptive *k*-smoothing). These two approaches share the same underlying exploratory model, effect measure (relative risk), and reference population (total study population) but lend themselves to different modes of visual summarization (shaded regions vs. image and surface plots). A third method (*k*-smoothing–derived polycircles) summarized the results of adaptive *k*-smoothing using an exclusive reference group.

Regionalization: multiscale grids. We divided a 37 × 37 km area encompassing the five Upper Cape Cod towns into three grids (large, medium, and small scales). Figure 2 shows the 16 (4 × 4), 64 (8 × 8), and 256 (16 × 16) grid cells, whose respective dimensions are 9.3, 4.6, and 2.3 km. We then coded subject residency into a dichotomous (yes/no) variable for each grid cell. A "yes" value represents "ever having lived within" a grid cell. We counted each subject only once per grid cell, but some may have been counted in more than one grid cell because the data included historical residences; we did not count residency after the diagnosis year for cases, or after the index year for controls. We

Table 1. Enrollment of breast cancer cases and controls.

Enrollment category	Cases	HCFA controls	Deceased controls	Random-digit dial controls
Selected	334	611	918	2,236
Excluded				
Never located or contacted	33	21	97	456
Ineligible	6	53	27	1,531
Physician or subject refusal	30	73	71	65
Interviewed	265	464	723	184

Table 2. Selection of breast cancer cases and controls for analysis.

Selection category	Cases	Controls
Interviewed	265	1,371
Selected controls		
Within age and vital status strata of cases	NA	763
Exclusions		
Moved to Cape Cod after index year	NA	46
Incomplete residential history	7	31
Analyzed	258	686

NA, not analyzed; these categories apply only to controls.

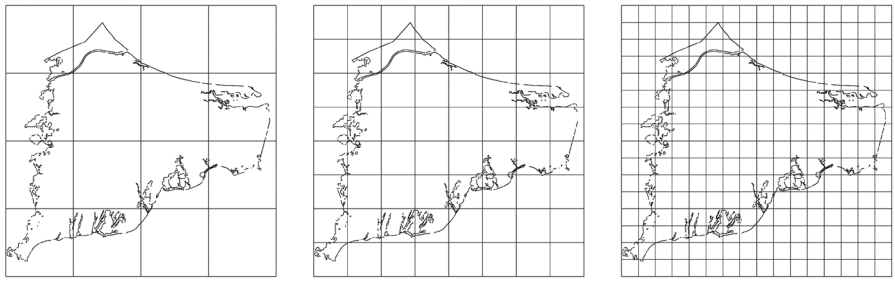
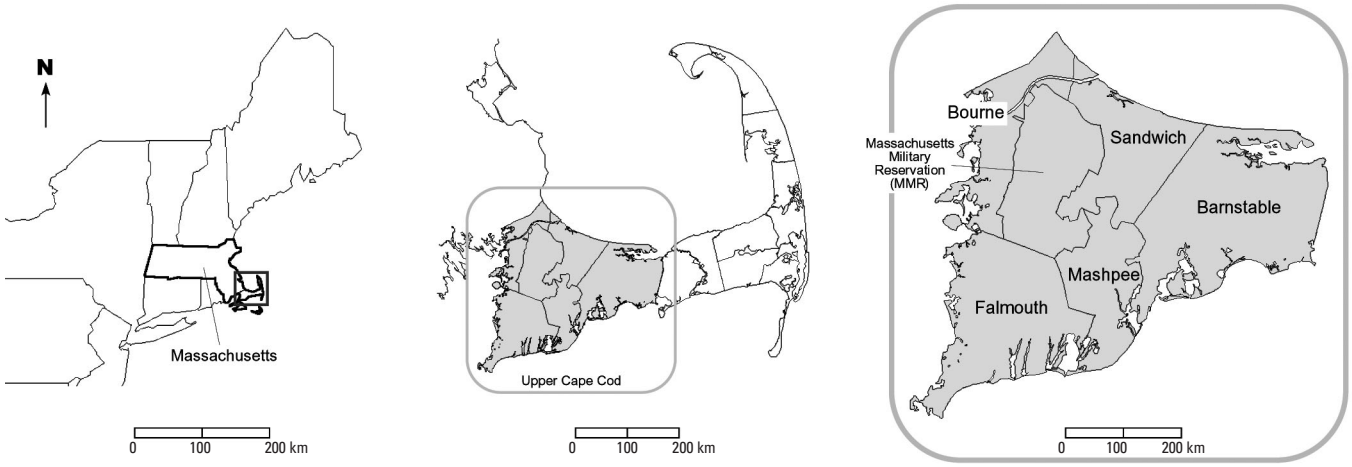


Figure 2. Multiscale grids.



Northeast United States | Cape Cod, Massachusetts | Upper Cape Cod

Figure 1. Location of study area.

removed from analysis grid cells containing fewer than three cases or fewer than three controls to facilitate adjusted analyses and curb the influence of unstable estimates.

Regionalization: adaptive* k*-smoothing. A second method for regionalizing residential location used a simultaneous smoothing and stabilization parameter. Adaptive *k*-smoothing defines regions using each subject's residential location as a focal point of a circle (Figure 3). The circle is just large enough to contain a prespecified number, *k*, of control subjects. Thus, the size of the circle used to regionalize the data is determined by the *k*th nearest neighboring control subject to the case or control focal subject. We used three values for *k*—50, 30, and 10 controls—in separate analyses, to investigate how stability and scale alter the spatial distribution of association. As in the grid cell method, we defined regional membership in terms of residency history. We considered a case or control a member of the circular region surrounding the focal residence if the subject ever resided within its bounds. In contrast to the multiscale grids, the circular regions defined by the set of *k* nearest neighbors are overlapping and act as data smoothers. The algorithm also adapts to differences in underlying population density, allowing spatially high-resolution (small circle) analysis where data are dense and lower resolution (large circle) analysis where data are sparse.

Measure of effect and reference population. We used a rate ratio measure of effect, computing relative disease incidence measures that compare subgroups with the study population as a whole.

For every region (s_i) created by a grid cell or adaptively defined circle, we calculated a case-to-control ratio: the number of cases of disease divided by the number of controls. This ratio is conventionally called a "disease odds." Rothman and Greenland (*16*) term this ratio a "pseudo-rate," because it is similar to a rate. The pseudo-rate comprises the number of cases divided by a *sample* of the person-time giving rise to them, whereas a rate would comprise the number of cases divided by the total person-time. We divided each regional group's pseudo-rate by the pseudo-rate of the entire study population. This odds ratio (OR) estimates the incidence rate ratio of within-region ($s_1 \ldots s_n$) incidence to total study population incidence. We used the entire study population as the reference population to derive comparable rate ratio estimates using the various regionalizations, and also to ensure a sufficiently large reference group to provide stable quantification of the associations between and among disease status and potentially confounding attributes.

Adjusted ORs and potential confounders. We used multiple logistic regression models to control simultaneously for potentially confounding individual attributes. For every subregion ($s_1 \ldots s_n$), we constructed a regression model containing an indicator variable for the regional membership attribute (e.g., $s_1 = 1$ for subjects residing in s_1; $s_1 = 0$ for the reference population), additional variables for the selection attributes (age at diagnosis or index year, vital status at time of interview), and potential confounders: family history of breast cancer in a first-degree relative (mother or sister), age at first live birth or stillbirth (by age group vs. nulliparous), and prior history of breast cancer or benign breast disease. We used the modeled coefficient of the regional membership variable to compute an adjusted OR.

Visual summarization. For the grid ORs, we divided crude ORs > 1 into quartiles and then color-shaded each of the upper three quartiles of the grids. This rendered a choropleth, or map of regions shaded by value. We used the same cut points to create choropleths of adjusted ORs. We replicated the process for the three grid scales.

We referenced crude and adjusted ORs from adaptive *k*-smoothing as point locations (the residential address of each subject). For these results, we used continuous modes of

Table 3. Distribution of selected characteristics of breast cancer cases and controls.

Characteristic	Percent cases (n = 258)	Percent controls (n = 686)
Caucasian	98.4	96.8
Age at diagnosis or index year (years)		
1–49	12.0	7.1
50–59	14.0	10.2
60–69	31.4	33.1
70–79	26.4	29.4
≥ 80	16.3	20.1
Educational level at least 12 years	83.1	82.7
Alive at interview	67.4	55.1
Age at first live birth or stillbirth (years)		
< 30	55.4	61.9
≥ 30	14.0	13.3
Nulliparous	30.6	24.8
Prior breast cancer or benign breast disease[a]	14.5	23.2
Family history of breast cancer[b]	19.2	8.8
Postmenopausal at diagnosis or index year	88.0	92.0
Religion		
Roman Catholic or Protestant	92.6	94.0
Jewish	3.5	3.1
Other/no response	3.9	2.9

[a]Before current diagnosis for cases. [b]Mother and/or sister.

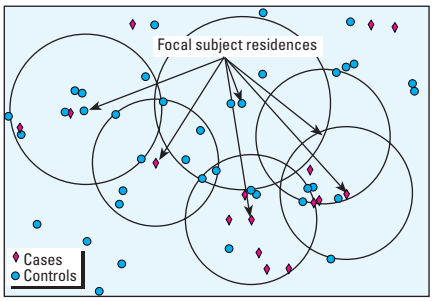


Figure 3. Example of circles created for adaptive *k*-smoothing: selection of circles drawn just large enough to include a prespecified number (*k*) of controls.

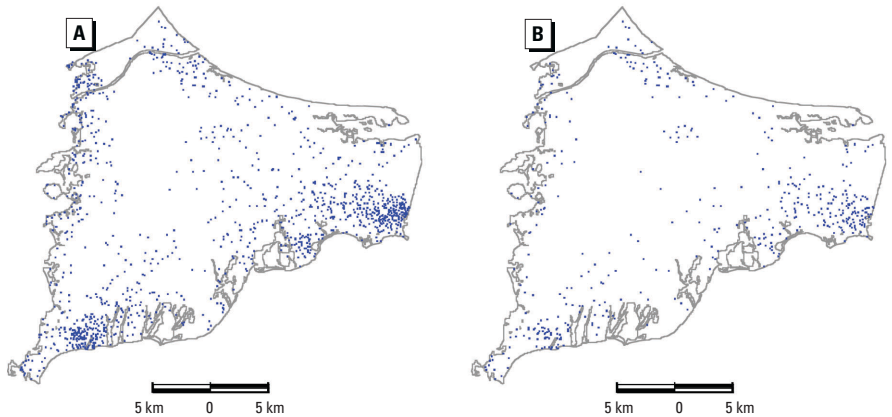


Figure 4. Distribution of control (*A*) and case (*B*) residences. Each dot indicates one subject residence, randomly placed within 1.2 km grid cells (not shown).

representation—image and surface plots—after interpolating between point locations. We accomplished interpolation by kriging, a method that fills in values between points by means of a sophisticated data-averaging algorithm (*17*). Weights are derived by calculating the dependence of variability on the degree of separation between observed points. The inverse variance estimates then become the weights in a weighted average of data values. We generated image and surface plots from the continuous longitude (x), latitude (y), and kriged OR (z) data. Surface plots represent the OR values as height above the x–y plane, in a three-dimensional perspective. Image plots code the OR values on a continuous rainbow color scale (violet to red, signifying low to high values), creating something akin to a choropleth but with finer spatial and OR resolution. We masked out the region falling outside the boundaries of the five-town study area of the image plots.

***Polycircles derived from* k*-smoothing*.** For each *k* parameter (50, 30, and 10 controls), we selected focal subjects whose *k*-smoothed OR was in the upper 2.5% of the distribution of adjusted ORs. This selection defined a set of circles—some overlapping and some not. We combined overlapping circles to create fixed-boundary regions, termed "polycircles." Because these polycircles were derived from the highest ORs obtained from *k*-smoothing, they constitute a collection of "hot spots."

We then computed two different types of ORs. First, a crude OR compared the pseudo-rate of subjects who ever lived within a polycircle to the pseudo-rate of subjects who never lived within any polycircle. Second, we estimated a multiregion adjusted OR by a multivariate logistic regression model that included a set of indicator variables denoting whether a subject had ever resided within each polycircle, plus the same set of selection variables and potential confounders as in the adjusted analyses for multiscale grids and *k*-smoothing circles. We used the modeled coefficient of the regional membership variable to estimate the adjusted OR for each polycircular region. We computed 95% confidence intervals (CIs) without adjustment for multiple comparisons (*18*). The reference population for these analyses is the subpopulation that only lived outside the polycircular regions.

Results

The women in this study population were mostly elderly, Caucasian, educated at the high school level or beyond, and postmenopausal (Table 3). As expected, a higher percentage of cases than controls reported a family history of breast cancer, nulliparity, and late age at first birth. However, a somewhat lower percentage of cases than controls reported a prior history of breast cancer and benign breast disease.

We found 451 Upper Cape residences from the 258 cases, for an average of 1.7 residences per case, and 1,111 control residences from 686 controls, for an average of 1.6 Upper Cape residences per control. Figure 4 shows dot density plots of control residences and case residences. Subject residences are spread along the coastlines, are absent from the central-west region (occupied by the Massachusetts Military Reservation), and otherwise tend toward the major town centers: Falmouth in the southwest and Hyannis (part of Barnstable) in the central-east. The most readily apparent difference in relative density of residences, comparing cases and controls, occurs in an area just slightly north of center, where a group of case residences is surrounded by an area free of case residences; the same region is scattered relatively uniformly with control residences.

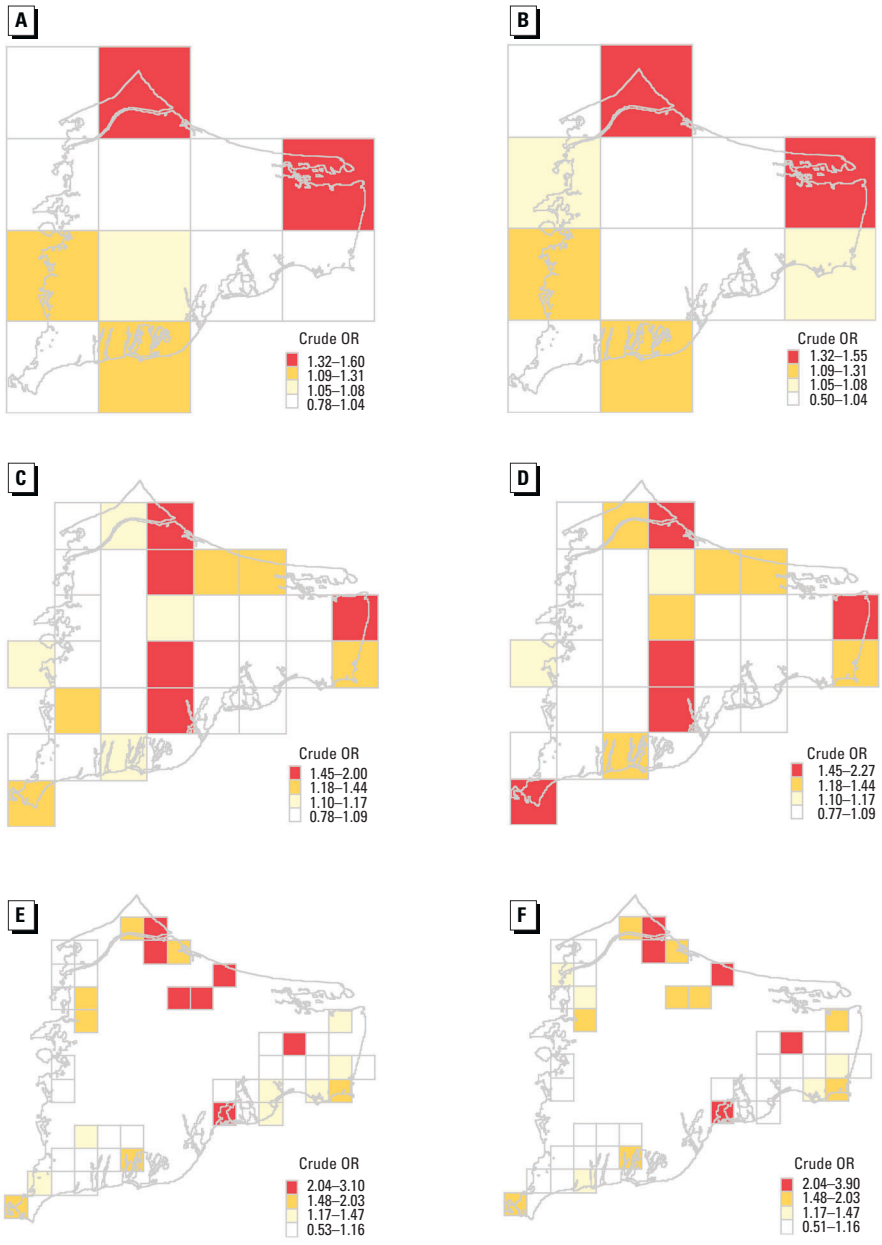


Figure 5. Crude and adjusted ORs by grid cells. All ORs are based on the number of subjects having lived within each grid cell, relative to the total number of cases (n = 258) and controls (n = 686) analyzed. ORs were adjusted for age, vital status, family history of breast cancer, age at first live birth or stillbirth, personal history of prior breast cancer or benign breast disease. (*A*) Crude ORs by 9.3 km grid cells. (*B*) Adjusted ORs by 9.3 km grid cells. (*C*) Crude ORs by 4.6 km grid cells. (*D*) Adjusted ORs by 4.6 km grid cells. (*E*) Crude ORs by 2.3 km grid cells. (*F*) Adjusted ORs by 2.3 km grid cells.

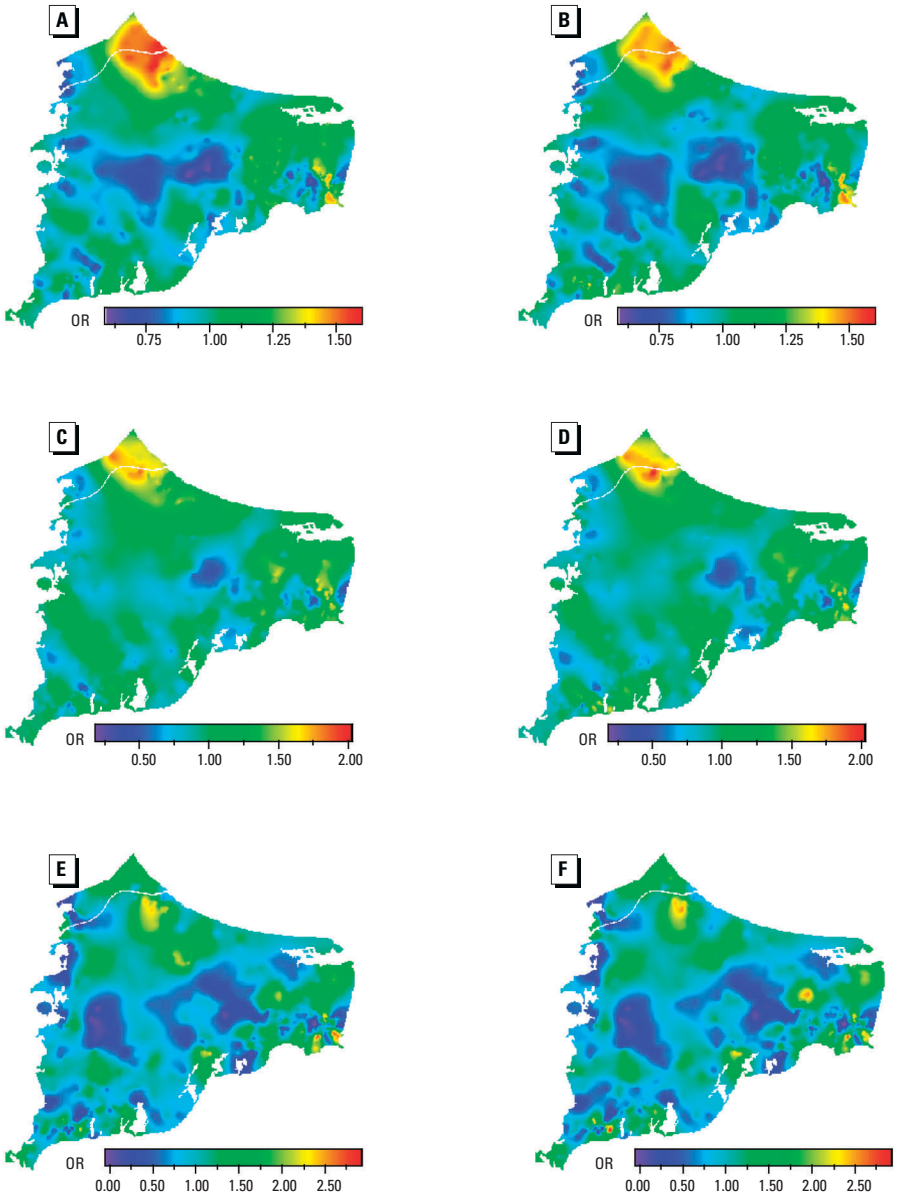


Figure 6. *k*-Smoothed crude and adjusted ORs. (*A*) *k*-Smoothed crude ORs (*k* = 50). (*B*) *k*-Smoothed adjusted ORs (*k* = 50). (*C*) *k*-Smoothed crude ORs (*k* = 30). (*D*) *k*-Smoothed adjusted ORs (*k* = 30). (*E*) *k*-Smoothed crude ORs (*k* = 10). (*F*) *k*-Smoothed adjusted ORs (*k* = 10).

Figure 5 shows choropleths of crude and adjusted ORs obtained at three different grid scales. The OR scale is specific to each grid scale, shading the three highest quartiles above OR = 1.0. Adjusted results (Figure 5B,D,F), shown to the right of each of the crude results (Figure 5A,C,E), use the same cut points as the crude plots. The limits of the highest and lowest categories have been extended as necessary to include the highest and lowest adjusted ORs. With decreasing geographic scale (top to bottom), data within each grid cell diminish and the variability of the ORs increases.

In general, although crude and adjusted analyses at each scale are very similar, patterns differ across scales. Large-scale areas associated with excess incidence tend to resolve into smaller-scale hot spots, as in the north, but large-scale grids with little or no excess can resolve into small-scale grids with elevated ORs, as in the east-central area. Some large-scale grids are not represented at smaller scales, because of a paucity of data, but most of the large-scale grids in the upper two quartiles of OR magnitude have some representation at smaller scales. The adjusted results show only subtle differences when compared with the crude results. A few grid regions shift between the two highest OR categories. The most consistent areas of excess are found straddling Sandwich and Bourne, to the northwest, and in Barnstable, to the east (see Figure 1 for town boundaries).

Figure 6 shows the pattern of crude and adjusted ORs obtained via *k*-smoothing. The OR scale differs for each *k* parameter (50, 30, and 10 controls, top to bottom, respectively) but is the same for crude and adjusted plots (left and right, respectively). At the largest *k* parameter of 50 controls, a large hot spot is apparent in the northwest area of the study region (straddling the border between Sandwich and Bourne), with localized peaks in the east (Figure 6A). Adjusting the crude values depresses the northwest peaks somewhat (OR ~1.5), elevates the eastern peaks (OR ~1.5), and elevates peaks in

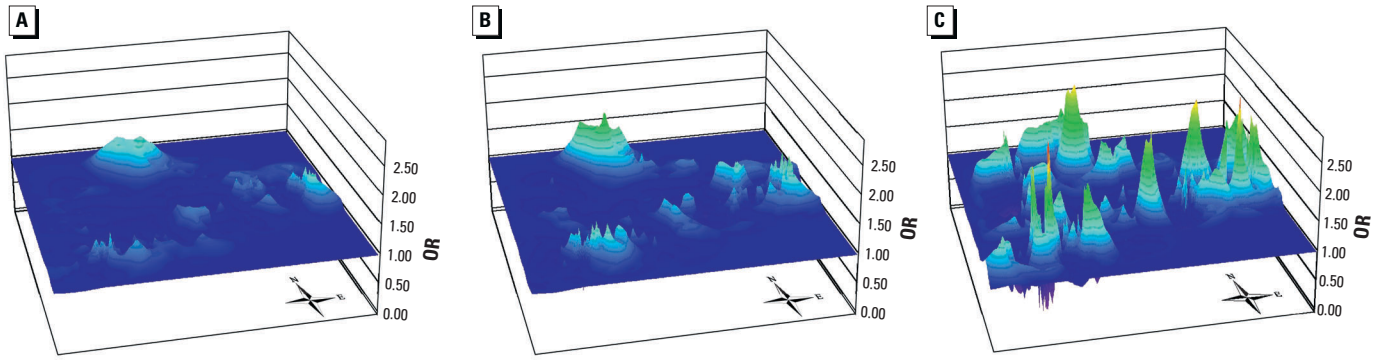


Figure 7. *k*-Smoothed adjusted ORs. (*A*) *k*-Smoothed adjusted ORs (*k* = 50). (*B*) *k*-Smoothed adjusted ORs (*k* = 30). (*C*) *k*-Smoothed adjusted ORs (*k* = 10).

the southwest (OR ~1.4) (Figure 6B). Reducing the *k* parameter to 30 controls presents a similar pattern of elevations and depressions (Figure 6C,D). However, some new peaks become apparent in a northwesterly direction from the eastern peaks, and one of the newly evident peaks persists after adjustment (OR ~1.5; Figure 6D). Reducing the adaptive smoothing parameter further, to 10 controls, yields still more variation (Figure 6E,F). In addition to the peaks seen in Figure 6A–D, peaks now appear in both the south-central (OR ~2.3) and east-central (OR ~2.1) portions of the study area. Adjustment had little qualitative effect on the results for this smoothing parameter in terms of either location or magnitude. The magnitude of the adjusted ORs ranges from 0 to 3.2. Placing the ORs on the same vertical scale gives the sequence of surface plots in Figure 7, which graphically shows the relation between *k*-value and spatial smoothing. (Because the surface plot wireframe contains averaged values, it imposes an additional degree of smoothing compared with the image plots.)

Selecting adjusted ORs among the highest 2.5 percentile for each *k* parameter gives a set of adaptively sized circles (Figure 8). Figure 9A–C displays the polycircles derived from these results. Table 4 summarizes the numbers of cases and controls ever having lived within each of the polycircular regions, and the related crude and adjusted ORs. Regional membership in a polycircle is defined nonexclusively as subjects who ever lived within the polycircle. The reference group is defined exclusively as subjects who never resided in any of the polycircles. The reference group is thereby a fixed subpopulation of 138 cases and 453 controls. For each *k*-specific polycircle analysis, some subjects belong to neither the reference group nor the *k*-specific polycircles. For instance, in the $k = 50$ analysis, Table 4, section $k = 50$ does not tabulate subjects who did not have a residence in the $k = 50$ polycircles and who did have a residence within the $k = 30$ or $k = 10$ polycircles. These subjects are in neither the reference group nor the $k = 50$ polycircles' population. The numbers of subjects shown in each section in Table 4 therefore sum up to less than the total number of cases and controls.

Residential membership in the three $k = 50$ polycircles (Figure 9A) is associated with a relative risk of 1.7–1.9, after adjustment (Table 4, top). Residential membership in either of the two $k = 30$ polycircles (Figure 9B) is associated with an approximate 2.0-fold relative risk, after adjustment (Table 4, middle), whereas adjusted ORs for the $k = 10$ polycircles range from 1.6 (95% CI, 0.8–3.2) to 3.1 (95% CI, 1.3–7.2). These relative risk estimates are relative to those subjects who never lived in any polycircle.

Discussion

We began the exploration of case and control residency distribution by making visual comparisons of dot maps. Such a comparison was problematic because compensating visually for the greater number of controls and the irregular boundary edges was difficult. These distributions may also be skewed representations of the numbers of persons in case and control groups, because they plot the history of residence (e.g., a clump of multiple residences may belong to a single case). This suggested the need for regionally grouped incidence rate ratio estimates.

Crude and adjusted ORs. We can interpret the calculated ORs as ratios of residence-within-region incidence to total study population incidence. The comparison is conservative for two reasons. First, the reference population is diluted by the subgroup of interest. Second, the baseline (OR = 1)

Figure 8. Adaptive circles with adjusted ORs in the upper 2.5 percentile. (*A*) $k = 50$ circles. (*B*) $k = 30$ circles. (*C*) $k = 10$ circles.

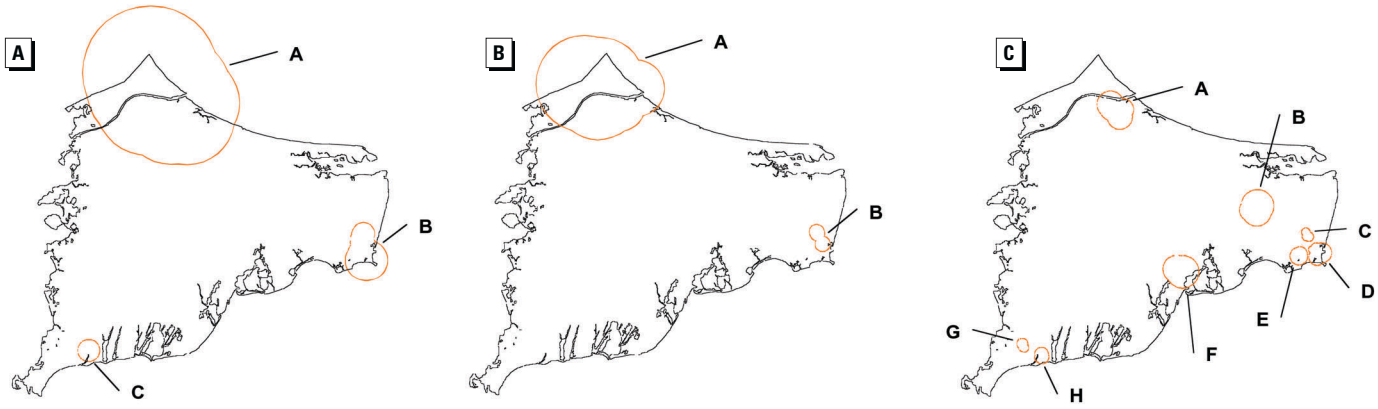


Figure 9. Polycircles for adaptively smoothed, adjusted analyses. (*A*) $k = 50$ polycircles. (*B*) $k = 30$ polycircles. (*C*) $k = 10$ polycircles. Letters next to polycircles correspond to data and results in Table 4.

comprises a population known to have an elevated incidence of breast cancer as a whole, compared with statewide, age-standardized incidence (*1,19*).

We considered several alternatives in choosing a reference group and method of multivariate adjustment. Dividing the study population into many subgroups and defining one of them as a reference category would produce less precise estimation the subgroups are numerous and any reference group would be small (*20*). This procedure also requires removing a subgroup from analysis. Adjustment using indirect standardization (standardized incidence ratios) is an alternative that would yield noncomparable estimates when the target populations differ in their distributions of the standardizing variable or weights (*21*). Furthermore, standardization will be invalid if the number of expected cases in any stratum is small (*21*).

Our resolution to the issue of adjustment was to extend the strategy employed for crude estimation: Use the entire study population as reference group. We could then obtain adjusted ORs via a logistic regression model that included a set of potentially confounding variables. The adjusted ORs can be thought of as comparing within-group incidence of disease (defined by residential location) with disease incidence of the entire study population, controlling for a set of attributes independently associated with the disease. As in the crude analysis, this choice of reference population produces conservative incidence rate ratio estimates but artificially narrowed precision estimates (not reported), because the total number of study subjects is artificially increased (total plus within-region population). These considerations led us to devise a follow-on analysis of polycircles that uses an exclusive reference group.

Regional membership. The grid analysis divided the study area into equal-sized grid cells at three scale levels, producing a set of crude and adjusted ORs for each scale. The observed associations between residential location and breast cancer incidence may be contingent on the specific set of grid boundaries. For example, a group of neighboring cases that, if taken together, would represent a stable excess could nonetheless be split into negligibly small subgroups by the grid boundaries. Also, areas of equal size frequently contain unequal populations and are not comparable in terms of precision. Although the issue is to some extent unavoidable with any set of nonoverlapping boundaries, it could be addressed in future investigations by simultaneously varying the scale, shape, and boundaries of grid units.

The method of adaptive *k*-smoothing avoids some of the drawbacks related to measuring ORs using grid cells. By stipulating a fixed number of controls (*k*) for the ratio of cases to controls for every adaptive circle, relative risk estimates of comparable magnitude are also of comparable numerical stability. In principle, applying adaptively sized, overlapping circles produces a smoother map of incidence rate ratios while preserving small-scale estimates where the underlying numbers are ample. Because the rate ratio estimates are spatially referenced to a large number of point locations, they can be interpolated and used with continuous modes of graphic representation such as image and surface plots. These visualization tools allow easy identification of high and low values. On the other hand, interpolated summarization of the adaptively smoothed results gives the impression that the study population is evenly distributed, even though it is not. Consequently, plateaus in the interpolated graphics do not necessarily represent areas with uniform level of incidence rate ratios. A plateau may instead represent a sparsely populated area, or even an unpopulated area, where distant relative risk estimates are similar because of the application of large circles. The spatial extent of features on adaptively smoothed maps needs to be interpreted cautiously.

Both the grid and *k*-smoothing methods, as we have implemented them, use a nonexclusive reference group, the total study population. This choice of reference group gives conservative incidence rate ratio estimates and inflated precision estimates (not reported). The bias toward the null is minor for small subpopulations and does not change rank order relations of rate ratio estimates within the study population. Assessing statistical stability, however, is a general issue that needs to be addressed. Numbers of cases and controls could have been documented for the grid analysis, but this would have been cumbersome to interpret. Adaptive *k*-smoothing has the advantage of obtaining estimates of similar stability across the study area for a given *k* parameter. However, adaptive *k*-smoothing does not allow a straightforward specification of the numbers of cases and controls, because the underlying circular units are overlapping and not visible in the graphical summaries. Neither method is designed to adjust for residency migration within the study area, so possible residency correlation between subregions is not accounted for.

These considerations led us to construct a follow-on analysis that summarizes and extends the results of adaptive *k*-smoothing. By cutting off a selection of peak values obtained from the adjusted *k*-smoothing analyses, and then joining any overlapping circles, we defined polycircular "hot spot" regions. This discrete set of polycircles offers several useful features: *a*) specific identification of the areas associated with the most elevated incidence rates, *b*) enumeration of how many cases and controls have resided therein, *c*) calculation of incidence rates relative to an exclusive reference group, and *d*) estimation of confidence intervals. The main disadvantage of the polycircle analysis is that it represents just one cutoff (in this case, the upper 2.5 percentile). However, areas that would be included by increasing the cutoff percentile can be inferred from the *k*-smoothing plots, which represent all results.

In the polycircle analysis, we computed confidence intervals without adjustments for multiple comparisons or the data-driven procedure of selecting peak results. The reported confidence intervals are based only on the observed data, using conventional single-inference methods. Also, each confidence interval is ascribed to a polycircle for which there is in fact only one data sample. We

Table 4. Crude and adjusted ORs for polycircular regions.

Region	No. cases	No. controls	Crude OR	Adjusted OR (95% CI)
k = 50				
A	33	63	1.7	1.7 (1.0–2.7)
B	47	93	1.7	1.8 (1.2–2.7)
C	23	51	1.5	1.9 (1.1–3.2)
Reference[a]	138	453	1	1
k = 30				
A	28	48	1.9	1.9 (1.1–3.2)
B	33	52	2.1	2.2 (1.3–3.5)
Reference[a]	138	453	1	1
k = 10				
A	14	16	2.9	3.0 (1.4–6.6)
B	9	13	2.3	2.2 (0.8–5.5)
C	15	19	2.6	2.4 (1.1–5.0)
D	17	25	2.2	1.6 (0.8–3.2)
E	11	12	3.0	2.6 (1.0–6.3)
F	9	10	3.0	2.9 (1.1–7.8)
G	9	13	2.3	2.3 (0.9–5.9)
H	11	14	2.6	3.1 (1.3–7.2)
Reference[a]	138	453	1	1

[a]Never lived in any polycircle for *k* = 10, 30, or 50.

believe these intervals offer a more interpretable assessment of statistical precision than do intervals that would incorporate somewhat arbitrary aspects of analysis such as the total number of comparisons made or the percentile cutoff of peak values. Adjustments for multiple comparisons would increase the type II (false negative) error rate and assume a universal null hypothesis (*18*). Neither of these traits is justified in an exploratory analysis, where one assumes that some associations are truly non-null, and the goal is to uncover any evidence leading to more specific identification of the factors involved.

Interpretation of results. Breast cancer incidence in this study population was not uniformly distributed with respect to residential location. Residential location within three subregions—in the northwest, southwest, and east—was associated with increased incidence of disease relative to the study population as a whole. These hot spots were consistently observable at different spatial scales and using both fixed and adaptive spatial boundaries. Adjusting for age and other individual risk factors had only minor influence on the overall spatial distribution of incidence rate ratios.

To assess the possibility that the observed associations represent one or more underlying, geographically situated environmental risk factors, we must consider alternative explanations. First, the observed spatial associations may be a reflection of geographic confounding: any confluence of nonenvironmental factors associated with residential location, independently associated with the disease outcome, that thereby produces a spatial association.

We used a multivariate logistic regression analysis to control for confounding by the selection variables (age and vital status), and a set of variables representing characteristics known to be most strongly associated with breast cancer. Although some of the overall variability of ORs decreased, and specific ORs increased or decreased compared with the crude results, neither the overall pattern nor specific localities exhibiting excess changed appreciably. However, confounding by unknown individual risk factors or mismeasurement of the selected potential confounding variables could account for the spatial associations we have highlighted and may have positively or negatively influenced other parts of the spatial distribution of association.

An alternative explanation for the observed spatial associations is that the initial sampling selection of controls was spatially biased. We know that nonenrolled members of the target population were similar to enrolled subjects in demographic characteristics; we do not know if they were also similar in residency distribution. Other errors, such as misplacement of mapped residences, could also haphazardly create positive and negative bias in the observed spatial associations.

Ultimately, the influence of factors unknown to investigators cannot be assessed directly. The possibility that an observed result reflects a "chance" association can only be judged indirectly, in terms of numerical stability and statistical precision estimates. Many of the relative risk estimates presented here are numerically unstable, because of the aim of assessing small regions with consequently small populations. Numerical stability can be enhanced at the expense of spatial precision. Our approach has been to look at multiple geographic scales and to place greater emphasis on regional associations revealed at more than one scale. A small-scale association may be dismissed as measurement error, but a larger, more stable calculation is more difficult to discount.

Although it is impossible to exclude the possibility that the results we have highlighted are due to "chance," the associations are stable enough to argue against this conclusion as an explanation for the hot spots.

Conclusions

Our aim was to use case–control data from the Upper Cape Cod Cancer Incidence Study to identify small-scale hot spots of breast cancer incidence that may lead to causal explanations in further investigations. The findings suggest several directions for further research. One path is to explore subject interviews for individual characteristics and nonenvironmental factors that might distinguish cases from their controls who resided in the observed hot spot regions. Another complementary task is to analyze the extent to which previously investigated environmental exposures (*1,3,6,7*) contribute to the observed spatial patterns. The task then will be to explore these extensive environmental history data in a Geographic Information System (*19*) for particular features of the hot spot regions with the potential for human exposure (e.g., past pesticide application). These tasks will assist in generating causal hypotheses that can be assessed in a forthcoming case–control study. The addition of more years of incidence data could also reveal areas of residence that are persistently associated with excess incidence.

References and Notes

1. Aschengrau A, Ozonoff D. Upper Cape Cod Cancer Incidence Study. Final Report. Boston:Massachusetts Department of Public Health, 1992.
2. Aschengrau A, Ozonoff D, Paulu C, Coogan P, Vezina R, Heeren T, Zhang Y. Cancer risk and tetrachloroethylene-contaminated drinking water in Massachusetts. Arch Environ Health 48:284–292 (1993).
3. Aschengrau A, Paulu C, Ozonoff D. Tetrachloroethylene-contaminated drinking water and the risk of breast cancer. Environ Health Perspect 106(suppl 4):947–953 (1998).
4. Paulu C, Aschengrau A, Ozonoff D. Tetrachloroethylene-contaminated drinking water in Massachusetts and the risk of colon-rectum, lung, and other cancers. Environ Health Perspect 107:265–271 (1999).
5. Webler T, Brown HS. Exposure to tetrachloroethylene via contaminated drinking water pipes in Massachusetts: a predictive model. Arch Environ Health 48:293–297 (1993).
6. Aschengrau A, Ozonoff D, Coogan P, Vezina R, Heeren T, Zhang Y. Cancer risk and residential proximity to cranberry cultivation in Massachusetts. Am J Public Health 86:1289–1296 (1996).
7. Ozonoff D, Aschengrau A, Coogan P. Cancer in the vicinity of a Department of Defense superfund site in Massachusetts. Toxicol Ind Health 10:119–141 (1994).
8. Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast cancer clusters in the northeast United States: a geographic analysis. Am J Epidemiol 146:161–170 (1997).
9. Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. Am J Epidemiol 132:S136–143 (1990).
10. Openshaw S, Craft AW, Charlton M, Birch JM. Investigation of leukaemia clusters by use of a Geographical Analysis Machine. Lancet 1:272–273 (1988).
11. Cislaghi C, Biggeri A, Braga M, Lagazio C, Marchi M. Exploratory tools for disease mapping in geographical epidemiology. Stat Med 14:2363–2381 (1995).
12. Kafadar K. Geographic trends in prostate cancer mortality: an application of spatial smoothers and the need for adjustment. Ann Epidemiol 7:35–45 (1997).
13. Rushton G, Lolonis P. Exploratory spatial analysis of birth defect rates in an urban population. Stat Med 15:717–726 (1996).
14. Bithell JF. An application of density estimation to geographical epidemiology. Stat Med 9:691–701 (1990).
15. Webster R, Oliver MA, Muir KR, Mann JR. Kriging the local risk of a rare disease from a register of diagnoses. Geogr Anal 26:168–184 (1994).
16. Rothman KJ, Greenland S. Case-control studies. In: Modern Epidemiology (Rothman K, Greenland S, eds). Philadelphia:Lippincott, 1998;93–114.
17. Fortner B. The Data Handbook. Champaign, IL:Spyglass, Inc., 1992.
18. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1:43–46 (1990).
19. Brody JG, Rudel R, Maxwell NI, Swedis SR. Mapping out a search for environmental causes of breast cancer. Public Health Rep 111:494–507 (1996).
20. Greenland S. Analysis of polytomous exposures and outcomes. In: Modern Epidemiology (Rothman K, Greenland S, eds). Philadelphia:Lippincott, 1998;301–328.
21. Rothman KJ, Greenland S. Introduction to stratified analysis. In: Modern Epidemiology (Rothman K, Greenland S, eds). Philadelphia:Lippincott, 1998;253–279.