Spatial Analysis of Case-Control Data Using Generalized Additive Models

T. Webster, V. Vieira, J. Weinberg, A. Aschengrau

Departments of Environmental Health, Biostatistics, Epidemiology

Boston University School of Public Health

Presented at the EUROHEIS Symposium, 30-31 March 2003, Östersund Sweden

Correspondence:

Dr. Thomas F. Webster

Department of Environmental Health (T2E)

Boston University School of Public Health

715 Albany St.

Boston, MA 02130 USA

twebster@bu.edu

Disease registries typically record residence at time of diagnosis and contain information on only a few covariates. Maps relying on registry data are potentially subject to spatial confounding—the spatial clustering of risk factors—and exposure misclassification. Maps based on suitably conducted case-control studies can solve these problems, but methods for mapping such data have received relatively little attention (e.g., references 1-5).

METHODS

When cases and controls are appropriately sampled from the population of a geographic area, the case/control ratio (disease odds) in a sub-area should be proportional to the disease incidence rate in that sub-area. We estimate local disease odds via non-parametric regression. Generalized additive models (GAMs) provide an elegant approach, allowing simultaneous smoothing and adjustment for covariates, selection of optimum bandwidth (degree of smoothing), and hypothesis testing (6). Our method is related to that of Kelsall and Diggle(2) with some important differences. We assume the model:

$$\text{logit}[p(x_1,x_2)] = S(x_1,x_2) + \gamma'\mathbf{z} \qquad (1)$$

where the left-hand side is the disease log odds at location (x_1,x_2), $\mathbf{z}$ is a vector of covariates, γ is a vector of usual regression parameters, and $S(x_1,x_2)$ is a bivariate smoothing function. We employ loess, a smoother that is adaptive to data of varying density such as population. Loess combines advantages of nearest neighbor and fixed kernel approaches.

We used Splus and Arcview for our analysis. 1) We select an optimum bandwidth by minimizing the Akaike's Information Criterion and use equation (1) to estimate the crude (no covariates) or adjusted log odds at each location on a grid.

We use the deviance of model (1) vs. the reduced ("null") model without the smoothing term as a global statistic for the importance of location. Splus computes a p-value for this statistic, but deviance is only approximately chi square distributed for GAMs (6). 2) We permute the locations of the cases and controls and rerun the GAM using the same bandwidth as before. We repeat this process many times, generating: a) the distribution of the deviance under the null hypothesis of a flat map, b) the distributions of the log odds at every grid point under the null hypothesis. We use the former to estimate the p-value for the deviance test and the latter to construct a p-value at each location and a pointwise test under the null hypothesis at the 0.05 level. Alternatively, one can use bootstrapping to construct variance bands, a non-parametric counterpart to confidence intervals (7). 3) We convert from log odds to odds ratios using the whole study area as the reference: we divide the odds at each point by the "null" odds produced by equation (1) without the smoothing term. For crude models, this is equivalent to dividing by the ratio of the total number of cases to the total number of controls. 4) We overlay the grid of odds ratios on a map of the study area. We use a blue-red color spectrum and linear scale for the odds ratio to make results easier for a general audience to understand. Tests with synthetic data indicated that the method can properly adjust for covariates (7).

We investigated the association between residential history and colorectal, lung, and breast cancer on Upper Cape Cod, Massachusetts (USA) using data from two case-control studies of cancer with extensive data on covariates and a forty year residential history (8,9). Eligible subjects had to be permanent residents of the study area for at least six months during 1983-93 for breast cancer, 1983-86 for lung and colorectal cancer. Controls were assigned "reference years" in a manner that reflects the distribution of diagnosis years among cases. We examined latency by

restricting inclusion to the residences occupied by subjects a specified number of years prior to diagnosis or reference year.

RESULTS

Little or no change was seen in the maps of the three cancers after adjusting for a number of potential confounders, indicating that the risk factors that we analyzed did not account for the spatial patterns of disease. The maps for colorectal cancer were quite flat; the global test for the importance of location had p-values of 0.31 for no latency and 0.78 for 15 years of latency. Increasing latency increased the p-values of the location term for both lung cancer (0.16 for no latency, 0.002 for15 years) and breast cancer (0.22 for no latency, 0.006 for 15 years, 0.002 for 20 years). Relatively similar p-values were calculated using the global statistic of Kelsall and Diggle (2). Hot and cold spots tended to become more pronounced as we increased latency in the lung and breast cancer maps. Figure 1 shows adjusted maps for the three cancers.

Overlaying maps of odds ratios with maps of pollution sources can generate hypotheses about exposure. Caution is need because many geographic features may overlap. A significant lung cancer “hot spot” was located near the Massachusetts Military Reservation. Earlier research had found a modest increased risk of lung cancer within 3 km of gun and mortar training sites on the military base (10). The three significant breast cancer hot spots were located near ground water plumes. However, subjects may not have been supplied with water impacted by these plumes as many subjects used public water.

Our analyses included all subject residences allowed by the latency assumptions. Since some subjects had more than one residence, they would

appear in more than one location. To see if this might bias our breast cancer map with twenty years of latency, we restricted our analysis to the residence of longest duration. The map was relatively unchanged except that the areas of pointwise significance were reduced in size.

DISCUSSION

Generalized additive models provide a useful tool for mapping population-based case-control data. In our preliminary analyses, lung and breast cancer maps of the Upper Cape Cod area became more pronounced with increasing latency, while the maps for colorectal cancer were flat. We are currently exploring potential sources of bias including residual spatial confounding. Since several areas of elevated risk are near the coast, edge effects must also be considered. Additional methods are needed for examining multiple residences.

ACKNOWLEDGEMENTS:

This project was funded by Superfund Basic Research Program grant 1P42ES 07381.

REFERENCES:

1. Bithell J. (1990). An application of density estimation to geographic epidemiology. *Statist Med* 9: 691-701.
2. Kelsall J and Diggle P (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Appl Statist* 47: 559-573.
3. Paulu C, Aschengrau A, Ozonoff D (2002). Exploring associations between residential location and breast cancer incidence in a case-control study. *Environ Health Perspect* 110: 471-8.
4. Vieira V, Webster T, Aschengrau A, Ozonoff D (2002). A Method for Spatial Analysis of Risk in a Population-Based Case-Control Study. *Intern J Hygiene Environ Health* 205: 115-120.
5. Sabel CE, Gatrell AC, Loytonen M, Maasilta P, Jokelainen M (2000). Modelling exposure opportunities: estimating relative risk for motor neurone disease in Finland. *Soc Sci Med* 50:1121-37.

6. Hastie T, Tibshirani R. *Generalized Additive Models*. Chapman & Hall/CRC, 1990
7. Webster T, Vieira V, Weinberg J, Aschengrau A, Ozonoff D (2002). A Method for Mapping Population-Based Case-Control Studies Using Generalized Additive Modeling. ISEE/ISEA, Vancouver BC, August . *Epidemiology* 13: S256.
8. Paulu C, Aschengrau A, Ozonoff D (1999). Tetrachloroethylene-contaminated drinking water in Massachusetts and the risk of colon-rectum, lung, and other cancers. *Environ Health Perspect* 107:265-71
9. Aschengrau A, Rogers S, Ozonoff D (2003). Perchloroethylene-contaminated drinking water and the risk of breast cancer: Additional results from Cape Cod Massachusetts. *Environ Health Perspect* 111(2):167-74
10. Ozonoff D, Aschengrau A, Coogan P (1994). Cancer in the vicinity of a Dept. of Defense superfund site in Massachusetts. *Toxicol.Indust.Health* 10: 119-141.

Figure 1. Adjusted odds ratio maps for colorectal, lung and breast cancer on Upper Cape Cod