A Spatial Scan Statistic for Survival Data

Lan Huang,[1] Martin Kulldorff,[2] David Gregorio[3]
[1]SRAB, SRP, DCCPS, NIH/NCI
6116 Executive Blvd., Rockville, MD 20852
[2]Department of Ambulatory Care and Prevention,
Harvard Medical School and Harvard Pilgrim Health Care,
Boston, MA 02215
[3]Department of Community Medicine,
University of Connecticut School of Medicine,
Farmington, CT 06030

Abstract

Spatial scan statistics with Bernoulli and Poisson models are commonly used for geographical disease surveillance and cluster detection. These models, suitable for count data, were not designed for continuous outcome data. We propose a spatial scan statistic based on an exponential model to be used for uncensored or censored continuous survival data. The power and sensitivity of the developed model are investigated through intensive simulations. The method performs well for different survival distribution functions including the exponential, gamma and log normal distributions. How to adjust the analysis for covariates is described in detail. The method is illustrated using survival data for men diagnosed with prostate cancer in Connecticut from 1984 to 1995.

Keywords: Geographical surveillance, Spatial scan statistic, Exponential model, Survival data, Censoring, Covariate adjustment

1 Introduction

Spatial and space-time scan statistics are commonly used for geographical disease cluster detection. For mortality and incidence data, a Poisson model is used when the number of cases is compared to an underlying population at risk derived from the census. Examples of such use include the study of geographical distribution of variant Creutzfldt-Jakob disease in Great Britain (Cousens, 2001) and the study of soft-tissue sarcoma and non-Hodgkin's lymphoma in France (Viel et al., 2000). By contrast, Bernoulli models are used for dichotomous variables, such as early/late disease stage, prevalence and treatment data. One example is the study of geographical differences in primary therapy for early-stage breast cancer in Connecticut (Gregorio et al., 2001) and another is the spatial variation of late detection of breast and colorectal cancer in Minnesota (Thomas and Carlin, 2003).

Survival is another health outcome for which geographical disease studies are of interest. For example, Karjalainen (1990) studied the geographical variation in cancer patient survival in Finland, and Jack et al. (2003) studied the geographical differences in lung cancer survival in Southeast England. Recent work on spatial survival analysis has utilized spatially-structured frailties in Cox-type regressions. Banerjee et al. (2003) explored frailty models for spatially correlated survival data, with application to infant mortality in Minnesota, Li and Ryan (2002) developed semiparametric frailty models for spatial survival data, and Banerjee and Carlin (2003) studied semiparametric spatio-temporal frailty models for survival data.

Here, our problem is to determine if there are geographical clusters of people with shorter than expected survival time that may reflect inadequate treatment, more aggressive disease or differential health practices. Similarly, locations characterized by longer than expected survival time may reveal treatment advances or favorable prognostic indicators among the population at risk. A common feature of survival data is the presence of censored observations arising when knowledge of an individual's life length is known only to a certain point of time. How to deal with censored data is well established in conventional survival analysis, whereas their appropriate incorporation in spatial analysis is uncertain.

A naïve approach to analyze geographical survival data is to define a cut point, divide the subjects into short and long survival groups and employ a Bernoulli based spatial scan statistic on the dichotomized data. Not only is it hard to choose the cut point, there is a loss of information when changing the continuous data into a 0/1 variable. Moreover, if there are censored observations, it is not clear how to dichotomize the data when the censoring times are prior to the cut-off value.

Here we propose a spatial scan statistic based on the exponential model, to analyze geographic variation in health events measured on a continuous scale, such as survival time and disease progression. The exponential distribution is often used for survival data and it is ideal for handling censored observations. One concern is the sensitivity of the exponential model to other survival time distributions. A robustness study is implemented and the results for cluster detection of survival time modeled according to true exponential, gamma or log normal distributions are examined.

A second consideration is how to adjust the analysis in order to account for possible confounding factors such as age, gender, race/ethnicity, and disease stage, which could bear upon finding shorter or longer survival times. Hence, a way to adjust for covariates is presented.

The Bernoulli and exponential models and their corresponding spatial scan statistics are described in sections 2 and 3 respectively. The robustness of the exponential spatial scan statistic for other distribution functions is evaluated in section 4. The advantage of the exponential model over the Bernoulli model is also investigated. A practical example, applying the exponential model to prostate cancer survival data in Connecticut is given in section 5. In section 6, we show how to adjust for covariates. The paper ends with a discussion in section 7.

2 Bernoulli Model

A naïve approach for geographical survival data is to dichotomize the continuous time to event data, and then use the existing Bernoulli based spatial scan statistic to study the spatial variation in survival. While not recommending such an approach, we describe and evaluate it in comparison to the exponential model proposed later in this paper.

The Bernoulli model has been described in detail by Kulldorff (1997). In summary, let G be the study area and let Z be any circular sub-area in G. Let p be the probability of "short survival" for a case living within the zone $Z \in G$, while the same probability for individuals outside the zone is q. The null hypothesis $H_0 : p = q$ is contrasted to the alternative hypothesis $H_a : p > q$. Under H_0, the outcome for any one individual is independent of those for the others.

Let c_Z be the total number of cases with short survival and n_Z the total number of subjects within area Z. Let C and N be the respective totals for the whole study area G.

The Bernoulli based spatial scan statistic is defined as (Kulldorff, 1997)

$$\lambda = \max_Z (\frac{c_Z}{n_Z})^{c_Z} (1 - \frac{c_Z}{n_Z})^{n_Z - c_Z} (\frac{C - c_Z}{N - n_Z})^{C - c_Z} (1 - \frac{C - c_Z}{N - n_Z})^{(N - n_Z) - (C - c_Z)} \times I(\frac{c_Z}{n_Z} > \frac{C - c_Z}{N - n_Z}). \quad (1)$$

Note that there are many overlapping Z's in G. One zone Z could be centered at any location of the diagnosed patients, and with gradually increasing radius until the zone includes 50% of the total number of patients in the study region G. Therefore, it could include some areas with zero population, such as lakes, oceans, forests and areas outside G but having common border with G. The zone $\hat{Z}$, which maximizes the likelihood in equation (1), is called the most likely cluster.

As there is no closed form for the distribution of λ, Monte Carlo hypothesis testing is used. M data sets are generated under H_0, where $p = q$, and λ is computed for each simulated data set. At the α-level, H_0 is rejected if the rank of the λ obtained from real data is among the $\alpha(M + 1)$ largest λ's (Dwass, 1957).

If there is interest in clusters with longer survival times, the classification of short and long survival time may be reversed. If there is simultaneous interest in both types of clusters, the indicator function in equation (1) is removed from the definition of λ.

3 Exponential Model

3.1 Test Statistic

For a spatial scan statistic with an exponential model, let the survival time for each individual inside zone Z be distributed according to the exponential distribution with mean θ_{in}, while the survival times for individuals outside Z be exponentially distributed with mean θ_{out}. The null hypothesis $H_0 : \theta_{in} = \theta_{out}$ for any Z is contrasted with the alternative $H_a : \theta_{in} < \theta_{out}$ for at least one Z when one wants to detect clusters with shorter survival, with $H_a : \theta_{in} > \theta_{out}$ for at least one Z when one wants to detect clusters with longer survival, and with $H_a : \theta_{in} \neq \theta_{out}$ for at least one Z when one wants to find clusters with either shorter or longer survival. Note that the zone Z could be any circle with different centroid and size in the whole study area G. Under the alternative hypothesis that θ_{in} and θ_{out} are different for different zones, while the parameter Z disappears under the null hypothesis.

Suppose that there are N individuals under study and that associated with the i^{th} individual is a lifetime T_i and a fixed censoring time L_i. For the time being, the T_i's are assumed to be independently and identically distributed (i.i.d.) with the exponential probability density function $f(T_i) = \frac{1}{\theta}e^{-T_i/\theta}$. The lifetime T_i of an individual will be observed only if $T_i \leq L_i$. If $T_i > L_i$, the survival time is censored considering a right censoring mechanism. Define the observed time $t_i = \min(T_i, L_i)$. Let $\delta_i = 1 \; if \; T_i \leq L_i$, and $\delta_i = 0 \; if \; T_i > L_i$, where δ_i indicates whether the lifetime T_i is censored or not. Define $r_{in} = \sum_{i \in Z} \delta_i$ (the number of non-censored individuals inside zone Z), and $r_{out} = \sum_{i \notin Z} \delta_i$ (the number of non-censored individuals outside zone Z). Let n_{in} and n_{out} be the total number of individuals inside and outside the zone Z respectively. The total number of individuals in G is $N = n_{in} + n_{out}$ and the total number of non-censored individuals is $R = r_{in} + r_{out}$.

The likelihood for an arbitrary zone Z can be expressed as

$$L(Z, \theta_{in}, \theta_{out}) = \prod_{i \in Z} \frac{1}{(\theta_{in})^{\delta_i}} e^{-\frac{T_i \delta_i}{\theta_{in}}} e^{-\frac{L_i(1-\delta_i)}{\theta_{in}}} \times \prod_{i \notin Z} \frac{1}{(\theta_{out})^{\delta_i}} e^{-\frac{T_i \delta_i}{\theta_{out}}} e^{-\frac{L_i(1-\delta_i)}{\theta_{out}}},$$

$$= \frac{1}{(\theta_{in})^{r_{in}}} e^{-\sum_{i \in Z} \frac{t_i}{\theta_{in}}} \frac{1}{(\theta_{out})^{r_{out}}} e^{-\sum_{i \notin Z} \frac{t_i}{\theta_{out}}}.$$

where $i \in Z$ indicates that the ith individual is located in zone Z.

The related likelihood ratio test statistic for a test with the alternative $\theta_{in} \neq \theta_{out}$ for at least one zone Z is

$$\lambda = \frac{\max_{Z, \theta_{in} \neq \theta_{out}} L(Z, \theta_{in}, \theta_{out})}{\max_{Z, \theta_{in} = \theta_{out}} L(Z, \theta_{in}, \theta_{out})} = \frac{L(\hat{Z})}{L_0},$$

where $\hat{Z}$ is the zone which maximizes $L(Z, \theta_{in}, \theta_{out})$ under H_a, and L_0 is the maximum of $L(Z, \theta_{in}, \theta_{out})$ under H_0 (i.e., $\theta_{in} = \theta_{out}$ for any Z). We use the set of circular zones centered at one of the patient locations. Unlike the general scan statistic for count data where each zone may have only a single individual, here each zone consists of at least two and at most $\frac{N}{2}$ individuals. $L(\hat{Z})$ and L_0 can be calculated according to the following equations. Given an arbitrary zone Z, the maximum likelihood estimates of θ_{in} and θ_{out} are $\hat{\theta}_{in} = \frac{r_{in}}{\sum_{i \in Z} t_i}$ and $\hat{\theta}_{out} = \frac{r_{out}}{\sum_{i \notin Z} t_i}$. We then have:

$$L(\hat{Z}) = \max_Z \frac{1}{(\hat{\theta}_{in})^{r_{in}}} e^{-\frac{\sum_{i \in Z} t_i}{\hat{\theta}_{in}}} \frac{1}{(\hat{\theta}_{out})^{r_{out}}} e^{-\frac{\sum_{i \notin Z} t_i}{\hat{\theta}_{out}}},$$

$$= \max_Z \left(\frac{r_{in}}{\sum_{i \in Z} t_i}\right)^{r_{in}} e^{-r_{in}} \left(\frac{r_{out}}{\sum_{i \notin Z} t_i}\right)^{r_{out}} e^{-r_{out}},$$

Likewise

$$L_0 = \frac{1}{(\hat{\theta}_G)^R} e^{-\frac{\sum_{i \in G} t_i}{\hat{\theta}_G}} = \left(\frac{R}{\sum_{i \in G} t_i}\right)^R e^{-R}.$$

The censoring information is incorporated into the likelihood functions L and L_0 through r_{in}, r_{out}, and R, which are functions of the censoring indicator δ. Note that L_0 depends only on

R, the total number of non-censored individuals, but not on the spatial distribution of those individuals. For the alternative $\theta_{in} \neq \theta_{out}$ we can now write the test statistic as:

$$\lambda = \frac{\max_Z (\frac{r_{in}}{\sum_{i \in Z} t_i})^{r_{in}} (\frac{r_{out}}{\sum_{i \notin Z} t_i})^{r_{out}}}{(\frac{R}{\sum_{i \in G} t_i})^{R}}.$$

For the alternative $\theta_{in} < \theta_{out}$, this function is multiplied by $I(\frac{r_{in}}{\sum_{i \in Z} t_i} < \frac{r_{out}}{\sum_{i \notin Z} t_i})$, and for the alternative $\theta_{in} > \theta_{out}$, it is multiplied by $I(\frac{r_{in}}{\sum_{i \in Z} t_i} > \frac{r_{out}}{\sum_{i \notin Z} t_i})$.

If there are no censored observations, then under the alternative $\theta_{in} \neq \theta_{out}$:

$$L(\hat{Z}) = \max_Z (\frac{n_{in}}{\sum_{i \in Z} t_i})^{n_{in}} e^{-n_{in}} (\frac{n_{out}}{\sum_{i \notin Z} t_i})^{n_{out}} e^{-n_{out}},$$

and

$$L_0 = (\frac{N}{\sum_{i \in G} t_i})^N e^{-N},$$

and hence,

$$\lambda = \frac{\max_Z (\frac{n_{in}}{\sum_{i \in Z} t_i})^{n_{in}} (\frac{n_{out}}{\sum_{i \notin Z} t_i})^{n_{out}}}{(\frac{N}{\sum_{i \in G} t_i})^{N}}.$$

3.2 Permutation Test Procedure

Statistical inference and hypothesis testings will be conducted based on the distribution of the test statistic λ. Unfortunately, as with most scan statistics (Glaz et al., 2001), we cannot find the distribution of the test statistic in closed analytical form. One common approach for scan statistics is to generate simulated data under the null hypothesis, but that is difficult in our case, since the distribution of survival times are unknown in terms of both the expected survival time and the censoring mechanism. Instead, we will condition on the observed set of survival times and censoring indicators, permuting these observed pairs $\{(t_i, \delta_i),\ i = 1, \cdots, N\}$ among the individual geographical coordinates. Note that the individual geographical coordinates are the fixed locations of the individuals observed in the study region G.

To obtain the exact distribution for λ, the test statistic must be calculated for all $n!$ rearrangement (permutations) of the n observations, for which the computational demand is very large even for small data sets. Therefore, instead of a complete permutation, a random selection of 9999 permutations is used. The test statistic λ is calculated for each permutation, and p-values are determined using Monte Carlo hypothesis testing as proposed by Dwass (1957). With 9999 replications, the null hypothesis is rejected at the 0.05 α-level if the value of the test statistic for the real observed data set is bigger than the 500^{th} highest values of the test statistic coming from the replications plus the observed one. The corresponding p-value is $R/(1+9999)$, where R is the rank of the test statistic value from the observed data among all 10000 test statistic values.

In addition to the most likely cluster, there are always secondary clusters that can be ranked according to their likelihood values. Rather than comparing the n^{th} highest likelihood in the real data set with the n^{th} highest likelihood in the random data sets, we compare it with the maximum likelihood in the random data sets. The interpretation of this is that statistically significant secondary clusters that do not overlap with a more likely cluster are capable of rejecting the null hypothesis on their own account, irrespectively of any other clusters in the observed data.

Note that we are using two different null hypotheses, the first being a subset of the second. Formally, the null hypothesis is stated as the survival times being exponentially distributed with $\theta_{in} = \theta_{out}$ for all Z. This formulation is needed to derive the test as a likelihood ratio test statistic. In the permutation step though, we only require that all the survival times are equally distributed irrespectively of their spatial location, and this weaker definition of the null hypothesis is sufficient to ensure valid statistical inference.

To model the survival times, we used the exponential distribution. Many survival times are not exponentially distributed, but may follow a gamma, log normal, Weibull or other distributions. What is important is that because of the permutation based test procedure (we are not generating the simulated data under null from exponential distribution, but randomly permuting the locations and the survival time/censoring attributes of the observations), the statistical inference is still valid and the correct α level maintained irrespectively of the true underlying survival distribution. The exponential distribution, or to be more precise, the exponential based likelihood function, simply assigns a weight to the different survival times. For the exponential model, this weight is based on the continuous survival and censoring times observed, as opposed to the 0/1 weights used by the Bernoulli model which is only based on whether the observation is above or below the chosen cut point. In essence, the exponential derivation provides a summary value for each potential cluster considered by the procedure and the maximum of these values indicates the most likely cluster. The real data may follow other distributions, but one can still calculate the exponential based summary value. While this value for non-exponentially distributed data is not as precise as it is for exponentially distributed data and no longer has an interpretation as a likelihood test statistic, this does not matter since we don't use any likelihood theory to evaluate the statistical significance.

In the next section, we evaluate how robust the exponential based spatial scan statistic is for various survival time distributions.

4 Power, Sensitivity and Positive Predictive Value

To test the performance of the proposed method, survival data were randomly generated for 610 individuals. Different data sets were generated from exponential, gamma, and log normal distribution with different means and variances. For the geographical locations, we used real data, consisting of the locations of home residence of men diagnosed with prostate cancer in Connecticut in 1984. For all simulated data, a true cluster of 47 individuals was created in Fairfield County with the centroid at (41.079N, 73.618W) and a radius of 8.65km. Individuals within this cluster had a survival time with lower mean than the rest of the state.

We simulated data sets either with censoring or without censoring. Here, we only consider data with right censoring and the censoring times vary randomly according to some probability distributions inside and outside the true cluster. We let L_i have the same distribution as T_i for those individuals outside of the true cluster. The percentage of individuals with censored survival times are then about 50%.

We created 10000 simulated data sets for each probability model under alternative hypothesis. For each of these simulated data sets, we generated 999 random permutations to obtain the p-values. For each model, the power is estimated as $\frac{\text{number of simulations with p-value }<0.05}{10000}$. To evaluate the precision of the detected clusters, we define the sensitivity to be the proportion of the individuals in the true cluster 'captured' by the detected cluster and the positive predictive value (PPV) to be the proportion of the individuals in the detected cluster belonging to the true cluster. We estimate the sensitivity by

$$\frac{1}{S}\sum_{s=1}^{S}\frac{\text{number of individuals in both true and detected clusters in } s^{th} \text{ simulation}}{\text{number of individuals in the true cluster in } s^{th} \text{ simulation}},$$

and the PPV by

$$\frac{1}{S}\sum_{s=1}^{S}\frac{\text{number of individuals in both true and detected clusters in } s^{th} \text{ simulation}}{\text{number of individuals in the detected cluster in } s^{th} \text{ simulation}},$$

where S is the total number of simulations.

The results are shown in Table 1. As expected, the power is larger when the difference in the mean survival times is greater, and also when the survival times have smaller variance. The power is not only high for exponential survival times, but also for the other survival time distributions. Note that for identical values of the mean difference and variance, the power is higher for log normal compared to exponential survival time, even though the scan statistic is based on the exponential likelihood. As expected, the power for censored data is lower than that for noncensored data since there is a loss of information.

We used sensitivity and PPV to investigate if the location of the detected cluster is close to the location of the true cluster, as shown in Figures 1, 2, and 3. The means and medians of the sensitivity and PPV are high (≥ 0.9) when the related power is high (> 0.9), and are around 0.5 to 0.9 when the related power is moderate. The variability of the sensitivity and PPV increase when the power is low. Note that when the power is high or moderate, the distribution of sensitivity and PPV are both left skewed, with the mean lower than the median. When the power is low, the distributions are right skewed. The sensitivity is always smaller than or equal to the PPV, which implies that the detected cluster tends to be somewhat smaller than the true cluster, but at the right location.

For comparison purpose, we evaluated the Bernoulli based scan statistic for noncensored survival data, using the median as the cut point to dichotomize the data. The Bernoulli model has lower power than the exponential model for exponential, gamma and lognormal data (Table 1). The exponential model is also better in terms of sensitivity and PPV (Figure 1 and 2). The sensitivity and PPV are lower for censored data than those for noncensored data using exponential model, which is consistent with the results observed from the power evaluations.

The exponential model does not work so well for normally distributed data. First, the exponential model requires that all observations are positive numbers, while normal data may be either negative or positive. Second, if the observations are positive, the power of the exponential based model still cannot compete with the Bernoulli model. For illustration, we simulated normal data with mean 105, 107 and variance 25, 49 inside the cluster and mean 110 and variance 100 outside cluster. For normal data with mean 105 inside, the power is 0.8968 for the exponential model, and 0.9971 for the Bernoulli model. For normal data with mean 107 inside cluster, the power is 0.1464 for the exponential model and 0.3783 for the Bernoulli model. Therefore, we do not recommend the use of the exponential model for normally distributed data or for data with other similar approximately symmetric distributions.

5 Prostate Cancer Survival in Connecticut

We illustrate the use of the exponential based scan statistic with data on prostate cancer survival in Connecticut. Between 1984 and 1995, the Connecticut Tumor registry recorded 22,612 invasive prostate cancer incidence cases (ICD-9-#185) among the state's population-at-risk (roughly 1.2 million males 20+ years old in 1990). Latitude-longitude coordinates for locations of home residence at time of diagnosis (+/- 50 meters) were successfully assigned to 20,598 records (91.1%). Follow-up of cases (1 to 5,867 days surviving following cancer diagnosis) was completed through December 2000. 1325 individuals with missing survival time information plus 212 cases that reported death on the day of diagnosis were excluded from the analysis. The 19,061 available records are analyzed in the following study. Of these, 10,308 records have complete time to death follow-up, while 8,753 records (45.9%) are right censored.

The most likely cluster with significantly reduced survival time was the vicinity of Waterbury in Western Central Connecticut (Area 1 in Figure 4 (left panel), p=0.001) for which men living within that locale when diagnosed with prostate cancer where estimated to have 29% higher risk to die with the disease during follow-up than similarly affected men living elsewhere. Secondary clusters with shorter survival were found around the city of Bridgeport in Southern Connecticut (Area 2, p = 0.001) and a localized area in Central Connecticut (Area 3; p=0.003). Area 3 is very small with radius 0.68 kilometers and only 36 patients, so it appears as a small dot on the map. Conversely, the most likely location of patients that experienced significantly longer survival time was in the North Central suburbs of greater Hartford (Area 4, p=0.001) where the risk of dying with prostate cancer was observed to be only 84% of those outside the area. Secondary locations with significantly longer survival times were found for two different suburban communities of Fairfield County (Areas 5 and 6, p=0.001 and 0.015, respectively). Note that both Area 2 and Area 5 extend outside the border of Connecticut. Half of Area 2 is located in the ocean with zero population. A small part of Area 5 is in New York State, but since the study region is the state of Connecticut, the area in New York State is also treated as if it had zero population.

We also evaluate survival using k-year survival probability, which is the probability of survival k-year after diagnosis with the disease. As shown in Table 3, the estimated 3-year, 5-year and 10-year survival probabilities in Area 1, 2, and 3 (detected short survival clusters) are all much less than those found outside. The survival probability was very low in Area 3, but since there are only 36 patients in this small cluster, the standard error for the estimated survival probability is big. Therefore, even with a very low survival probability, Area 3 is still not the most likely cluster of short survival.

These results call for further analysis for possible explanation of the observed clusters. It is worth noting, for example, that the patients average age at diagnosis in the short survival clusters are older than those found outside those locations (Table 2). This raises the question as to whether the detection of clusters is simply artifacts of the geographical variation in age distribution of prostate cancer patients. In order to find clusters not dependent on age (i.e., attributable to other contextual factors), it is necessary to adjust for age as a covariate. In the following section we describe how to do this.

6 Covariates Adjustment

6.1 Adjustment Procedure

In the previous sections, we discussed how to detect clusters using the spatial scan statistic based on the exponential and Bernoulli models respectively. Both models used the information about survival times, the locations and numbers of individuals, but no information about covariates. Potential covariates of interest include demographic variables, such as age, gender, race/ethnicity, socio-economic status, or education; behavioral variables, such as dietary; and physiological variables, such as tumor grade, tumor stage, histology, blood pressure, hemoglobin levels or heart rate. All these variables could be nuisance factors, which lead a non-homogenous study population. Here, we describe one way to adjust for covariates when applying the exponential based spatial scan statistic to search for clusters.

Consider a life time $T_i > 0$ and a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ associated with the lifetime T_i, where $i = 1, \cdots, N$. The vector $\mathbf{x}_i$ may include values of some known functions of quantitative variables and certain indicator functions of qualitative variables. We assume that $\mathbf{x}_i$ includes an intercept by taking $x_{i1} = 1$. Let $\boldsymbol{\beta}(= (\beta_1, \cdots, \beta_p)')$ denote the corresponding $p-\mathit{dimensional}$ vector of regression coefficients. We use the linear regression approach (Klein and Moeschberger, 1997) to model the covariate effects on survival. In this approach, the natural logarithm of the lifetime $Y_i = ln(T_i)$ is modeled. Since we propose a scan statistic based on an exponential model, we will do the covariate adjustment using an exponential regression model.

This model can be written as:

$$Y_i = \mathbf{x}_i^{'}\boldsymbol{\beta} + W_i,$$

with β_1 being the intercept. W_i is an error term, which follows an extreme value distribution with density $f_W(w) = e^w e^{-e^w}, -\infty < w < \infty$. Then, T_i has a density given by $f(t|\mathbf{x}) = \exp(-\mathbf{x}'\boldsymbol{\beta}) \exp\Bigg\{ - \exp(-\mathbf{x}'\boldsymbol{\beta})t \Bigg\}$, which is an exponential distribution with mean $\exp(\mathbf{x}'\boldsymbol{\beta})$.

From the model, we obtain the estimate of the parameters, $\hat{\boldsymbol{\beta}}$, using the observed data from the whole study area. The estimated mean of survival time for i^{th} subject is $\exp(\mathbf{x}_i^{'}\hat{\boldsymbol{\beta}})$. We then adjust the individual survival times based on the estimated mean so as to remove the effect of the covariates.

For example, if we only have age as a continuous covariate in the model with intercept, we obtain the $\hat{\beta}_2$ as negative value, which indicates that the survival time is shorter for older people. For the youngest person, the expected survival time is estimated as $\exp(\hat{\beta}_1 + \kappa\hat{\beta}_2)$, where $\kappa = min(x_{i2}), i = 1, \cdots, N$, N is the total number of observations. For i^{th} person with $x_{i2} > \kappa$, the expected survival will be lower than that for the youngest person. The ratio of the mean survival for old vs. young people is then $\exp(\hat{\beta}_2 \times (x_{i2} - \kappa))$. In order to adjust for age, we let $t_i^{adj} = t_i \times \exp(-\hat{\beta}_2 \times (x_{i2} - \kappa))$, where $i = 1, \ldots, N$. After the adjustment, we have $E(t_i^{adj}) = E(t_{i_0})$, i_0 indicates the person with the smallest age, which implies the expected survival times for all persons become the same as that for the youngest person.

For binary covariates or dummy variables, the lower level values can be selected to be the κ. In the presence of multiple covariates, the formula for adjustment can be extended as follows:

$$t_i^{adj} = t_i \times \exp\left\{ -\sum_{j=2}^{p} \hat{\beta}_j \times (x_{ij} - \kappa_j) \right\},$$

where $\kappa_j = min(x_{ij}), i = 1, \ldots, N, j = 1, \ldots, p$, $p-1$ is the number of covariates included in the model with intercept. Note that the κ_j could also be defined to be the maximum or mean of the j^{th} covariate, or other values. The main purpose is to make the expected survival time to be the same for subjects with different values of the covariates that need to be adjusted for.

The spatial models described in section 2 and 3 can then be employed on the adjusted survival times, $t_i^{adj}, i = 1, \ldots, N$.

6.2 Connecticut Prostate Cancer Data

According to the result from an exponential regression model applied to the survival data in the whole study region, age at diagnosis significantly affects the hazard rates ($P < 0.0001$); for a one-year increase in age at diagnosis the hazard rate is 1.067 times higher. Consequently, the relative survival time is shorter and the survival probability is lower. Results from the exponential based scan statistic on adjusted survival time data are presented in Tables 2 and 3, and in Figure 4 (right panel). These findings differ importantly from the unadjusted analysis. Clusters 3, 5 and 6 are no longer statistically significant when the age composition of patients across Connecticut is taken into account. On the other hand, the statistical significance of Clusters 1, 2, and 4 cannot be attributed to age at diagnosis as that factor is now controlled in the analysis. As yet unadjusted geographic factors pertaining to etiology (e.g.,lifestyle), disease (e.g., stage, grade), health status (comorbidities), health services (therapies, settings) and social conditions (deprivation, social support) merit consideration for explaining the three remaining clusters.

7 Discussion

The spatial scan statistics with the Bernoulli and Poisson models are useful in detecting clusters in spatial count data, but neither model is directly applicable to survival time data. The Bernoulli model could be employed with survival data after using a cut point to dichotomize the data into short and long survival, but dichotomization leads to a loss of information and a loss of power. The situation is even more problematic for the Bernoulli model when the data are censored, since it is not clear what to do with subjects with censoring times that are smaller than the cut point used. Another issue is the arbitrary selection of the cut point. For all these reasons, we do not recommend the use of the Bernoulli model for survival data, recommending instead the exponential model proposed in this article.

The exponential model does not lead to biased p-values associated with the most likely cluster even when the true survival times are not exponentially distributed. This is because the permutation of the observed coordinates and the survival times ensures that the correct α level is preserved, no matter how the survival times are distributed. The exponential model works well for censored and noncensored survival data, and for exponential, gamma, and log normal survival time distributions. The exponential model does not work well for all continuous data though, and we do not recommend its use for data that is approximately normally distributed. Further study should be done to find models that work well for such data.

The findings for prostate cancer survival are consistent with the observed variation in prostate cancer incidence in Connecticut during a roughly similar period (Gregorio et al., 2004). Areas that exhibited greater than expected incidence appear here to have experienced better than expected survival. This is most likely a consequence of the nature of prostate cancer, a screeningdetectable condition affecting men in advanced years of life. That is, circumstances conducive to good health and health care yield high numbers of cases (through vigilant screening) and these cases experience preferable outcomes (through early detection and/or treatment). Conversely, locations characterized by fewer than expected cases were found here to have poorer survival with the disease that may reflect other health disparities present in those locations, as well as inadequate detection and/or treatment of cases.

The method was used for purely spatial data in our paper. Scan statistics have also been developed for the space-time setting to detect clusters that exist in both space and time, in either a retrospective (Kulldorff et al., 1998) or prospective situation (Kulldorff, 2001, Kulldorff et al., 2005). The exponential model can be directly extended to such a setting by simply defining the zones Z as 3-dimensional cylinders rather than circles. It can also be extended to create a scan statistic with elliptical (Kulldorff et al., 2006) or other cluster shapes (Patil and Taillie, 2003, Tango and Takahashi, 2005, Duczmal and Assunção, 2005) if we suspect that the true shapes of the clusters are not approximately circular. The exponential model has recently been incorporated into the available SaTScan$^{\mathrm{TM}}$ software which can be downloaded from www.satscan.org.

8 Acknowledgments

This research was made possible through a Cooperative Agreement between the Centers for Disease Control and Prevention (CDC) and the Association of Teachers of Preventive Medicine (ATPM), award number U50/CCU300860 project number TS-0431; its contents are the responsibility of the authors and do not necessarily reflect the official views of the CDC or ATPM.

9 Reference

Banerjee, S., Carlin, B. P. (2003). Semiparametric spatio-temporal frailty modeling. *Environmetrics* **14,** 523-535.

Banerjee, S., Wall, M. M., Carlin, B. P., (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics* **4,** 123-142.

Chaput, E. K., Meek, J. I., and Heimer, R. (2002). Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut. *Emerging Infectious Diseases* **8(9),** 943-948.

Cousens, S., Smith, P., Ward, H., Everington, D., Knight, R., Zeidler, M., Stewart, G., Smith-Bathgate, E., Macleod, M., Mackenzzie, J., and Will, R. (2001). Geographical distribution of variant Creutzfeldt-Jakob disease in Great Britain. *Lancet* **357,** 1002-7.

Cox, D. R. (1972). Regression Models and life tables (with discussion). *Journal of the Royal Statistical Society* **B 34,** 187-220.

Duczmal, L., Assunção, R. (2005). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* **45,** 269-286.

Dwass M. (1957). Modified randomization tests for nonparametric hypothesis. *Annals of Mathematical Statistics* **28,** 181-187.

Glaz J., Naus J., and Wallenstein, S. (2001). *Scan Statistics.* Springer-Verlag New York, Inc.

Good, P. (1997). *Permutation Tests: a practical guide to resampling methods for testing hypotheses.* Springer-Verlag New York, Inc.

Gregorio, D. I., Kulldorff M., Barry L., and Samociuk H. (2002). Geographic differences in incidence rates and stage of disease for breast cancer in Connecticut,1991-1995. *International*

Journal of Cancer **100,** 194-198.

Gregorio, D. I., Kulldorff, M., Barry, L., Samocuik, H., and Zarfos, K. (2001). Geographical differences in primary therapy for early-stage breast cancer. *Annals of Surgical Oncology* **8(10),** 844-9.

Gregorio, D. I., Kulldorff, M., Sheehan, T. J., Samociuk, H. (2004). Geographic distribution of prostate cancer incidence in the era of PSA testing. *Urology* **63,** 78-82.

Hollander, M. (1999). *Nonparametric Statistical Methods.* John Wiley & Sons, Inc.

Hougard, P. (2000). *Analysis of Multivariate Survival Data.* Springer-Verlag New York, Inc.

Jack, R. H., Gulliford, M. C., Ferguson, J., and Moller, H. (2003). Geographical inequalities in lung cancer management and survival in South East England: evidence of variation in access to oncology services? *British Journal of Cancer* **88(7),** 1025-1031.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53,** 457-481.

Karjalainen, S. (1990). Geographical variation in cancer patient survival in Finland: chance, confounding, or effect of treatment? *Journal of Epidemiology and Community Health* **11(10),** 960-3

Klein, J. P. and Moeschberger, M. L.(1997). *Survival Analysis: techniques for censored and truncated data.* Springer-Verlag New York, Inc.

Kulldorff, M., Huang L., Pickle L., Duczmal L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine* in press.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R.M., Mostashari, F. (2005). A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine* **2,** 216-224.

Kulldorff, M. (2001) Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society* **A164,** 61-72.

Kulldorff, M., Athas, W., Feuer, E., Miller, B., and Key, C. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American Journal of Public Health* **88(9),** 1377-1380.

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

Lehmann, L. E. (1986). *Testing Statistical Hypotheses.* Springer-Verlag New York, Inc.

Li, Y. and Ryan, L. (2002). Modeling Spatial Survival Data Using Semiparametric Frailty Models. *Biometrics* **58,** 287-297.

Patil, G. P. and Taillie, C. (2003). Geographical and network surveillance via scan statistics for critical area detection. *Statistical Science* **18(4),** 457-465.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* **4,** 11.

Thomas, A. J. and Carlin, B. P. (2003). Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering. *Statistics in Medicine* **22(1),** 113-127.

Viel, J. F., Arveux, P., Baverel, J., and Cahn, J. Y. (2000). Soft-tissue sarcoma and non-Hodgkin's hymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels. *American Journal of Epidemiology* **152,** 13-19.

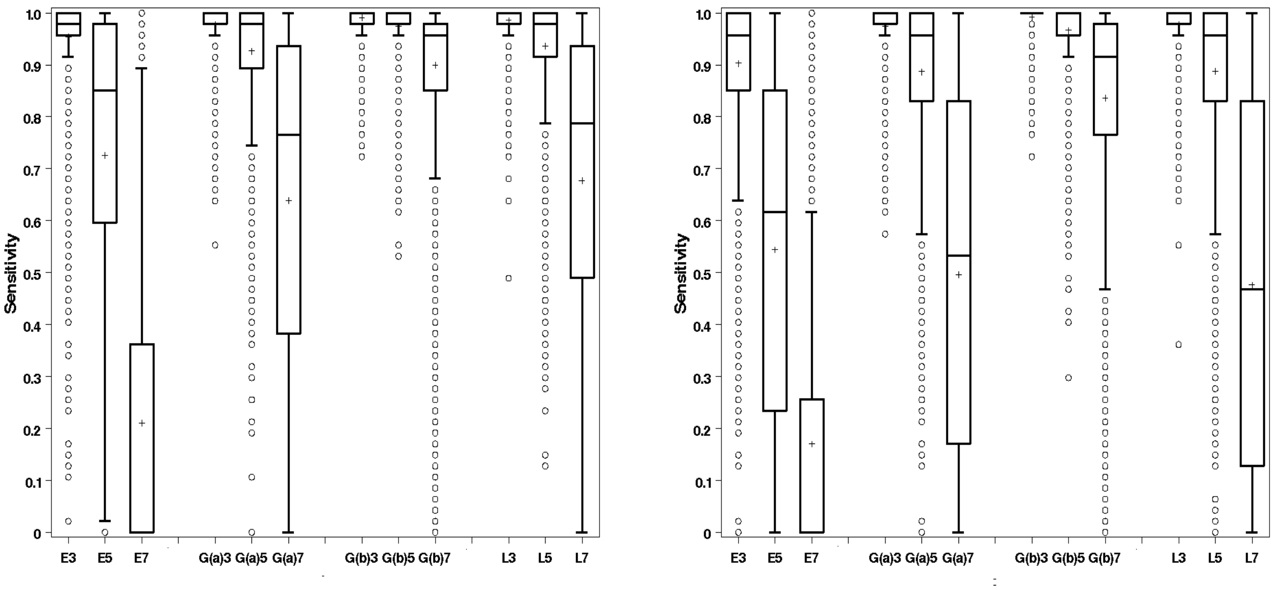

Figure 1: Side by side boxplot of Sensitivity from the **exponential** model (**left** panel) and **Bernoulli** model (**right** panel) on noncensored data. The labels on the x axis are consistent with those in Table 1. The + in the box indicates the mean, the line inside the box indicates the median, the upper border of the box is 75^{th} percentile, the lower border of the box is 25^{th} percentile. The interquartile range (IQR) is the difference between the 75^{th} percentile and the 25^{th} percentile. The outliers are shown as circles. Their values are defined as being above 75^{th} percentile+1.5(IQR) or below 25^{th} percentile-1.5(IQR).

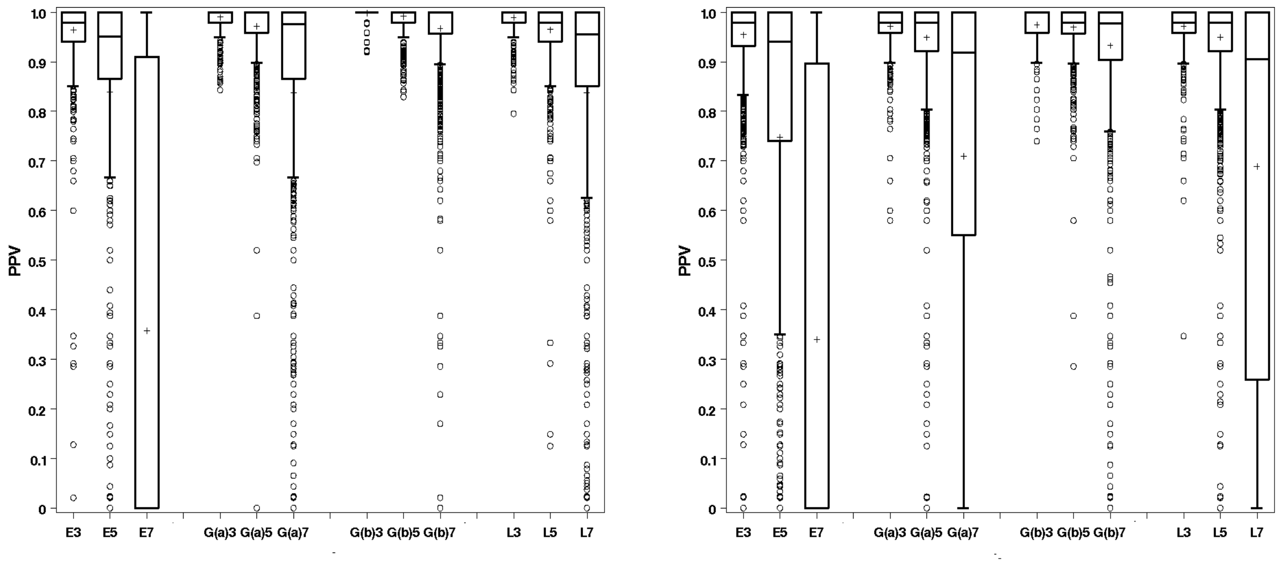

Figure 2: Side by side boxplot of PPV from the **exponential** model (**left** panel) and **Bernoulli** model (**right** panel) on noncensored data. The labels on the x axis are consistent with those in Table 1. All the symbols and lines are defined as in Figure 1.

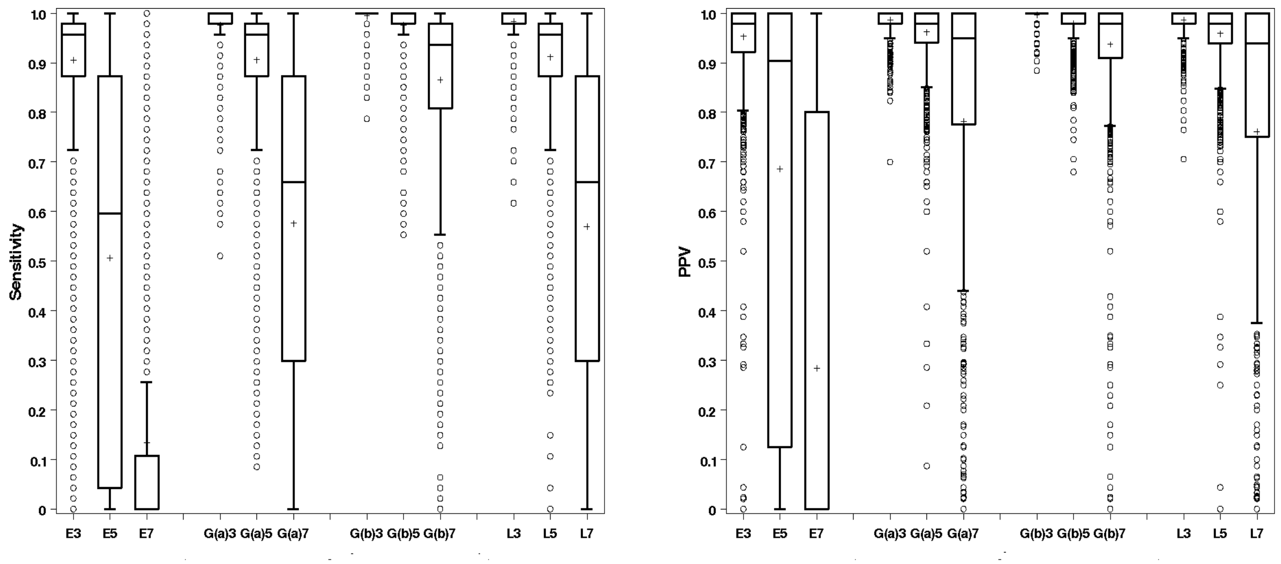

Figure 3: Side by side boxplot of Sensitivity (left panel) and PPV (right panel) from the exponential model on random censored data. The labels on the x axis are consistent with those in Table 1. All the symbols and lines are defined as in Figure 1.

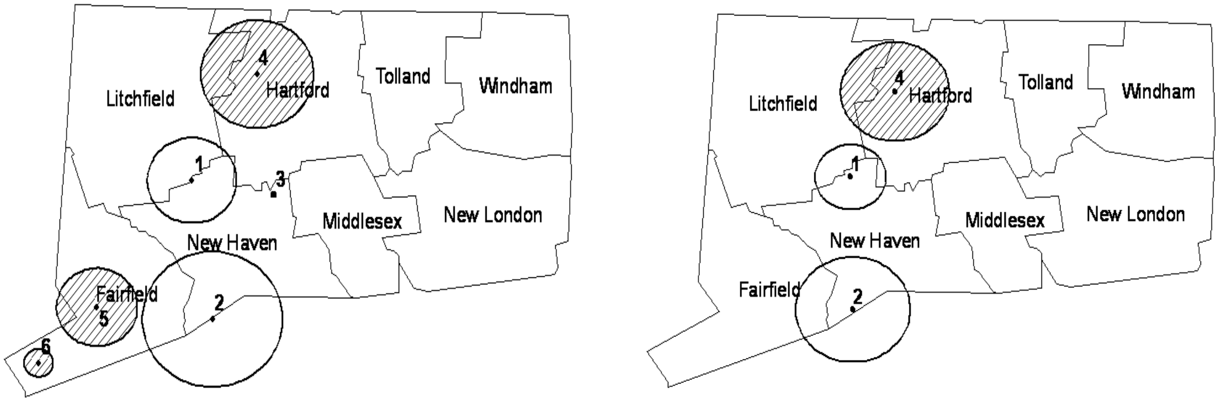

Figure 4: Significant clusters of short survival and long survival detected in Connecticut without age adjustment (left panel) and with age adjustment (right panel), with p-value smaller than 0.05. The circular regions with hatch are with longer survival time, and those without hatch are with shorter survival time.

clusters without age adjustment			clusters with age adjustments		
cluster	centroid	radius (KM)	cluster	centroid	radius (KM)
No.1	(41.577N, 73.085W)	12.69	No.1	(41.829N, 72.879W)	15.43
No.2	(41.206N, 73.015W)	20.06	No.2	(41.589N, 73.031W)	10.12
No.3	(41.538N, 72.806W)	0.68			
No.4	(41.861N, 72.862W)	16.00	No.4	(41.217N, 73.023W)	16.40
No.5	(41.238N, 73.409W)	11.57			
No.6	(41.089N, 73.608W)	4.16			

Table 1: Estimated power of the exponential and Bernoulli models applied to simulated data with different distributions. Outside the cluster, the mean for all the distributions is 10, the variance is 100.0 for exponential data, 31.6 for G(a) data, 10.0 for G(b) data, and 100.0 for log normal data.

Distribution of Simulated Data				Estimated Power	
Label	type	in cluster		Exponential	Bernoulli
		mean	variance	Model	Model
Noncensored data					
E3	exponential	3	9.0	0.9998	0.9917
E4		4	16.0	0.9554	0.8367
E5		5	25.0	0.6503	0.4940
E7		7	49.0	0.1090	0.1019
G(a)3	gamma	3	5.2	1.0000	1.0000
G(a)5		5	11.2	0.9994	0.9833
G(a)7		7	18.5	0.6432	0.4305
G(b)3	gamma	3	3.0	1.0000	1.0000
G(b)5		5	5.0	1.0000	1.0000
G(b)7		7	7.0	0.9992	0.9404
L3	log normal	3	9.0	1.0000	1.0000
L5		5	25.0	0.9998	0.9828
L7		7	49.0	0.6381	0.4012
Random censored data					
E3	exponential	3	9.0	0.9851	N/A
E4		4	16.0	0.7695	N/A
E5		5	25.0	0.3836	N/A
E7		7	49.0	0.0771	N/A
G(a)3	gamma	3	5.2	1.0000	N/A
G(a)5		5	11.2	0.9971	N/A
G(a)7		7	18.5	0.5238	N/A
G(b)3	gamma	3	3.0	1.0000	N/A
G(b)5		5	5.0	1.0000	N/A
G(b)7		7	7.0	0.9702	N/A
L3	log normal	3	9.0	1.0000	N/A
L5		5	25.0	0.9974	N/A
L7		7	49.0	0.4915	N/A

Table 2: The spatial scan statistic applied to prostate cancer data in Connecticut from 1984 to 1995 before and after adjusting the effect of age at diagnosis. The detected clusters and the corresponding p-values are presented. In the table, age is the average age at diagnosis, inside and outside the cluster respectively. RR is relative risk (= $\frac{(\#\text{ deaths in cluster A})/(\#\text{ individuals in cluster A})}{(\#\text{ deaths outside A})/(\#\text{ individuals outside A})}$). LLR is log-likelihood ratio.

cluster		in cluster		RR	LLR	P	age	age
		# deaths	# individuals				(in/out)	adjustment
short	1	646	938	1.29	41.88	0.001	72.1/71.4	no
survival	2	2154	3706	1.10	19.06	0.001	71.6/71.4	no
	3	33	36	1.70	16.13	0.003	72.2/71.4	no
long	4	661	1445	0.84	31.83	0.001	71.5/71.4	no
survival	5	200	529	0.69	22.24	0.001	72.0/71.4	no
	6	37	114	0.60	12.11	0.015	71.7/71.4	no
short	1	582	841	1.30	29.36	0.001	72.1/71.4	yes
survival	2	1662	2831	1.10	14.07	0.005	71.4/71.5	yes
long surv	4	704	1515	0.85	32.65	0.001	71.5/71.4	yes

Table 3: 3-year, 5-year, and 10-year survival probabilities estimated from Kaplan-Meier method (Kaplan and Meier, 1958) inside/outside the clusters shown in the maps with/without age adjustment.

cluster		cumulative survival probability (stderr)			adjustment
		3-year	5-year	10-year	
short	1	0.6856 (0.0152)	0.5362 (0.0163)	0.2371 (0.0180)	no
survival	2	0.7436 (0.0072)	0.6157 (0.0080)	0.3376 (0.0102)	no
	3	0.5000 (0.0833)	0.2222 (0.0693)	0.0864 (0.0522)	no
long	4	0.8348 (0.0098)	0.7427 (0.0116)	0.4903 (0.0163)	no
survival	5	0.8631 (0.0150)	0.7825 (0.0181)	0.5647 (0.0283)	no
	6	0.9192 (0.0258)	0.8555 (0.0334)	0.6831 (0.0510)	no
out clusters		0.7801 (0.0037)	0.6626 (0.0043)	0.3893 (0.0571)	no
short	1	0.6865 (0.0160)	0.5302 (0.0173)	0.2290 (0.0192)	yes
survival	2	0.7438 (0.0082)	0.6157 (0.0092)	0.3338 (0.0115)	yes
long surv	4	0.8307 (0.0096)	0.7391 (0.0113)	0.4862 (0.0159)	yes
out clusters		0.7804 (0.0035)	0.6637 (0.0040)	0.3922 (0.0054)	yes