Published in final edited form as:

Stat Med. 2008 November 10; 27(25): 5111–5142. doi:10.1002/sim.3342.

Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases

Lan Huang, Ph.D.¹, Linda W. Pickle², and Barnali Das²

¹Contractor, SRAB, SRP, DCCPS, National Cancer Institute, USA

²SRAB, SRP, DCCPS, National Cancer Institute, USA

Abstract

There have been articles on comparing methods for global clustering evaluation and cluster detection in disease surveillance, but power and sample size requirements have not been explored for spatially correlated data in this area. We are developing such requirements for tests of spatial clustering and cluster detection for regional cancer cases. We compared global clustering methods including Moran's I, Tango's and Besag-Newell's R statistics, and cluster detection methods including circular and elliptic spatial scan statistics (SaTScan), flexibly shaped spatial scan statistics (FSS), Turnbull's cluster evaluation permutation procedure (CEPP), local indicators of spatial association (LISA), and upper level set (ULS) scan statistics. We identified eight geographic patterns that are representative of patterns of mortality due to various types of cancer in the United States from 1998–2002. We then evaluated the selected spatial methods based on state- and county- level data simulated from these different spatial patterns in terms of geographic locations and relative risks, and varying sample sizes using the 2000 population in each county. The comparison provides insight into the performance of the spatial methods when applied to varying cancer count data in terms of power and precision of cluster detection.

Keywords

spatial statistic comparison; spatial clustering and cluster detection; cancer cases

1 Introduction

Cancer became a major public health focus in this country with the introduction of the National Cancer Act in the 1970s. An effort has been made to understand the progress in cancer treatment and trends in cancer incidence and mortality over time. There is also great interest in understanding the geographic distribution of cancer events, identifying the location of excess cancer cases, and evaluating the association between regional characteristics and cancer occurrence and death [1–4]. Research on disease surveillance has led to the development of many statistical tests for investigating the spatial variation of events, including methods for evaluating the tendency for global disease clustering and for assessing the location of the clusters [5]. Several questions then arise for cancer researchers. In order to understand the spatial patterns in cancer events, which method is more appropriate or more powerful? What are the required sample sizes and the degrees of excess events in terms of relative risk to achieve good power in detecting the spatial heterogeneity? Does the optimal method vary for different spatial pattern of cancer events? Can we provide

guidelines for the use of these statistical methods when applied to population-based cancer data? A few studies provide some understanding of the performance of tests for studying spatial randomness [10–12], but none of them focus on the comparison of cluster detection methods by the precision and sample size required to achieve good power for data with heterogeneous populations, realistic cancer patterns in the whole country, and a large number of geographic units.

Note that in addition to the methods discussed in [5], there is also a growing literature suggesting that cluster detection can be effectively carried out using Bayesian random effects models (Richardson et al 2004, Besag et al. 1991 and Knorr-Held and Rass 2000). Best, Richardson and Thomas (2005) provide a comprehensive review of those methods used for disease mapping within a Bayesian estimation paradigm and compare performance of representative models in this class. These methods use statistical spatial models to estimate relative risks by geographic units and researchers can then visually examine the maps for geographic patterns of varying risk. They do not directly identify an area as a statistically significant cluster as cluster detection methods do, or evaluate the global clustering pattern, so we do not include them in our study.

In this article, we intend to evaluate tests for global clustering and tests for cluster detection of typical spatial patterns of cancer. The population-based cancer data is usually aggregated count data instead of case-control data, because of the confidentiality of exact addresses. It is possible to obtain a random sample of controls from the population, but that is not of interest in our study. More than 100 global clustering methods are listed in Kulldorff's recent review [13]; we selected Moran's I [14], Besag Newell's R [15], and Tango's excess events test (EET) and maximized excess events test (MEET) [16–17], all of which have been widely used in the study of global spatial correlation. Many other methods have been proposed for global clustering evaluation, such as Geary's C [18], Oden's Ipop [19], Cuzick-Edward's k-Nearest Neighbors (k-NN) method [20], Swartz Entropy [21], and Whittemore test [22]. However, we decided not to examine these other tests for several reasons. Geary's C is similar to Moran's I, but centered at zero. One study suggested that Oden's Ipop has lower power than the standardized version of Moran's I [23]. The k-NN method was designed for case-control data, not for aggregated count data. Both Swartz' entropy and Whittemore's tests have been compared with Tango's MEET [12], and their performance was shown to be unstable. For cluster detection methods, we included SaTScan's circular version (SaTScanO) and elliptic version (SaTScanE) [24,25], Tango's flexibly shaped spatial scan statistic (FSS) [26], Turnbull et al.'s cluster evaluation permutation procedure (CEPP) [27], Patil and Taillie's ULS scan statistic [28], and the local indicators of spatial association (LISA) [29]. Two other well-known methods were also considered for cluster detection: the simulated annealing method [30] and Kelsall and Diggle's method [31]. However, in terms of power for detecting very irregularly-shaped clusters, SaTScanE is very competitive with the simulated annealing method [32], which was designed for irregularly-shaped cluster detection. Also, it is very computationally intensive to search for the irregularly-shaped clusters with the simulated annealing method as compared with the SaTScanE method. Kelsall and Diggle's method does not define any scan windows when searching for clusters in the study region. This method requires the location not only of cases, but of non-cases (controls) as well, to estimate the expected kernel surface; therefore, it requires more information and is not comparable with other methods. It is also very hard to obtain the geographic information for all non-cases in a large population. Because of these difficulties, we exclude both the simulated annealing method and Kelsall and Diggle's method from our study.

We provide a general description of selected spatial patterns of cancer mortality data in section 2, and discuss the tests for comparison in section 3. In section 4, an intensive

simulation study is implemented to compare the performance of the tests in terms of power and precision of the detected clusters. A discussion concludes the paper.

2 Spatial patterns of cancer

Among the quantile maps of the mortality rates in US, we selected eight cancer sites with global spatial patterns that seemed representative of patterns observed for mortality rates in 1998–2002 due to many different types of cancer. The generalized patterns for these sites, shown in Figure 1, formed the basis for our study. We examined the 3109 counties in 49 U.S. states, excluding Alaska and Hawaii (they are not contiguous with the continental United States and thus cannot be part of a broad, multi-state cluster of counties), and including the District of Columbia, both as a county and a state. The county relative risks range from 1.0 (lightest color) to $1+c$ (darkest color) with $c \geq 0$. We call the areas with relative risk larger than 1.0 cluster areas. Note that the clusters for kidney and lung male cancer sites have two different levels of relative risk ($1+c$ and $1+0.5c$). We can have varying relative risk values by assigning different c values. More details on c are described in section 3.5.

Information on the cluster areas regarding the number of counties in clusters, population, and urban-rural continuum (Beale) codes [33] appears in Table 1. The values of the Beale codes range from 1 to 9 with smaller values representing counties in more urban areas. The values in Table 1 are averaged over counties in the corresponding cluster areas. We called the clusters with small number of counties, or medium, and large number of counties if the numbers of counties inside clusters are < 400 , $400-600$, and $600+$, respectively. The patterns are also stratified into small, medium, and large population groups with average populations of $< 60,000$, $60,000-90,000$, and $90,000+$, respectively. The patterns with average Beale code values of $4.0-4.5$, $4.5-5.0$, and $5.0+$ represent clusters located in more urban, less urban, and least urban regions, respectively. The patterns are further characterized as follows.

- The spatial pattern for bladder cancer is typical of cancers where death occurred mostly in the highly populated (with the largest average population as 127,380) northeastern regions. This pattern has small number of counties, high population density, as well as clusters that are more urban.
- The pattern for cervical cancer has high rates in a band of counties in the south and southeast. This pattern has clusters with large number of counties, small population, and regions that are less urban.
- The pattern for colorectal cancer represents patterns of cancer death occurring in many disconnected clusters along the Appalachian mountain, with small number of counties, moderate population, and located in regions that are less urban.
- The pattern for kidney cancer has high rates in the middle of the country (least urban), with average Beale codes of 5.20. The number of counties included in the cluster areas is large for this site (698 counties), and the population is moderately large.
- The pattern for liver cancer has low rates in the north and high rates in the south covering areas with large population. The number of counties included is medium and located in the least urban regions.
- For prostate cancer, the cases occurred in the northwest and southeast, concentrated in the least urban areas. The clusters have medium number of counties and small population.

- The geographic pattern for lung cancer mortality is not the same for male and female, so these patterns are reported separately here. Female lung cancer cases are mostly in the northwest, with smaller clusters in other rural areas, and male lung cancer is concentrated in the areas around the southern Appalachian Mountains and the Gulf Coast. Female lung cancer clusters have only 324 counties and moderate population, and are located in the least urban regions. Male lung cancer clusters have 926 counties included, low population, and located in less urban regions.
- The maps show that the clusters for cervical, bladder, kidney, liver, and lung (male) sites are more concentrated in one area of the country, such as the mid-south for cervical, north east for bladder, central for kidney, south for liver, and mid-south for lung (male) sites. The clusters for colorectal, prostate, and lung (female) sites are less concentrated and are spread widely throughout the country.
- We do not calculate a measure to evaluate the parts of the clusters on the U.S. border that might reflect an edge effect of the cluster patterns. However, the maps show that the liver, cervix, and bladder sites have more clusters along the border; colorectal, kidney, and prostate have a moderate number; and female and male lung cancer sites have the least.

In this article, we simulate data based on the above spatial patterns and evaluate the performance of the methods (described in the next section) on the simulated data with varying spatial patterns, relative risks, and total numbers of cases (sample size).

3 Method of evaluation

3.1 General notation

Let n_+ be the total population observed in the study region G , which is the sum of the population in each geographic unit such as county or state (n_z , $z = 1, \dots, T$), where z indexes the geographic units in G , and T is the total number of geographic units in G (3109 for counties and 49 for states plus DC). Similarly, we use c_+ and c_z , $z = 1, \dots, T$ to denote the total number of cases in the whole country and in the geographic cells (z), respectively. Throughout this paper we use the generic term cell to mean a specific geographic area, which in our study is a county or state.

3.2 Test statistics for global clustering methods

Global clustering methods address global patterns of spatial correlation across the study area and are designed to detect the tendency of cases to cluster rather than to identify a particular collection of cases. We describe the three widely used methods that are investigated in this article: Moran's I , Tango's statistics, and Besag-Newell's R statistics.

3.2.1 Moran's I —Probably the most commonly used global index of spatial autocorrelation (overall clustering) is Moran's I [14]. This statistic measures the similarity of values in neighboring places from an overall mean value:

$$I = \left(\frac{1}{S^2} \right) \frac{\sum_{z=1}^T \sum_{z'=1}^T w_{zz'} (Y_z - \bar{Y})(Y_{z'} - \bar{Y})}{\sum_{z=1}^T \sum_{z'=1}^T w_{zz'}}$$

where $z \neq z'$. The weights $w_{zz'}$ could be general measures of influence or proximity but we used the adjacency definition, i.e., $w_{zz'} = 1$ if cells z and z' are adjacent, 0 otherwise; and

$s^2 = \frac{1}{T} \sum_{z=1}^T (Y_z - \bar{Y})^2$. The regular version of Moran's I tests the similarity of the numbers of cases (i.e., $Y_z = c_z$) in neighboring cells. The value of I is strongly positive if neighboring cells have similar values, either all mostly high or all mostly low. For this study we constructed the adjacency matrix, including elements $\{w_{zz'}\}$, using GeoDa (<http://www.geoda.uiuc.edu>) for all methods that required it.

The traditional calculation of Moran's I (regular version) as defined above does not account for population heterogeneity, so that its application to disease rates may result in an indication of clustering that is completely due to the spatial proximity of large population places, and not due to a cluster of high disease rates. Several alternative versions of Moran's I have been proposed to account for heterogeneous populations, for example by Oden [19], Waldhr [34], Walter [35], Assuncao and Reis [23], and Waller, Hill and Rudd [10]. We chose two of these alternatives for inclusion in our study:

- Replace the number of cases with the rate (i.e., $Y_z = \frac{c_z}{n_z}$) in the above formula of Moran's I (rate version), and
- Compare the observed number of cases with its expectation assuming constant relative risk, i.e.,

$$I^* = \frac{1}{\sum_{z', z=1}^T w_{zz'}} \sum_{z', z=1}^T w_{zz'} \left(\frac{Y_{z'} - rn_{z'}}{\sqrt{rn_{z'}}} \right) \left(\frac{Y_z - rn_z}{\sqrt{rn_z}} \right),$$

where $r = \frac{c_+}{n_+}$, the marginal rate over all locations. We refer to this as the normalized version of Moran's I.

3.2.2 Tango's statistics—Tango proposed an excess events test statistic for global clustering evaluation ([16]). For a given parameter θ , the statistic is defined as

$$eet(\theta) = \sum_{z=1}^T \sum_{z'=1}^T a_{zz'}(\theta) (c_z - n_z \frac{c_+}{n_+}) (c_{z'} - n_{z'} \frac{c_+}{n_+}),$$

where $a_{zz'}(\theta)$'s are the weights as a function of θ . Many weight functions have been proposed and compared in [36]. Here, we only consider two versions, adjacent neighbor weights (ADJ) and population density adjusted exponential weights (PDM). In the ADJ version, $a_{zz'} = w_{zz'}$ as discussed in Moran's I method. In the PDM version,

$a_{zz'}(\theta) = \exp \left\{ -4 \left(\frac{d_{zz'}}{\alpha_{zz'}} \right)^2 \right\}$. Here $\alpha_{zz'} = d_{zm_z}$, where $m_z = \max \{j : u(z, j) \leq \theta\}$, where $u(z, j)$ is the population size at location z and its j nearest neighbors. θ is a parameter defined by the user and can be viewed as a population measure for clustering. Usually, large θ is more sensitive to larger clusters and small θ is more sensitive to smaller clusters. For the ADJ version, there is no θ in the weight function, so the statistic is directly the EET. However, the PDM version of the test is sensitive to changes of the parameter θ , so in order to detect clustering irrespective of the geographic scale, Tango proposed the maximized excess events test (MEET) [17],

$$MEET = \min_{0 \leq \theta \leq U} P(EET(\theta) > eet(\theta)),$$

where we take values of θ for evaluation to be 5%, 10%, 15%, 45% and 50% of the total population. We set U to be 50% of total population.

3.2.3 Besag-Newell's R statistic—Besag and Newell has proposed a statistic to study clustering in rare diseases. We write it as

$$R = \sum_z c_z I \left[\sum_s \frac{e^{-u(z)} u(z)^s}{s!} > 0.95 \right],$$

where m is a fixed number of cases predetermined by users, $u(z) = \frac{c_+ \times u(z, f(z))}{n_+}$ and $u(z, f(z))$ is the population size at cell z and its $f(z)$ nearest neighbors. The statistic R is exactly the sum of the observed cases at selected cells (z 's). A cell is selected if a circular area centered at that cell has excess events. Specifically, a circle (Z) is first defined centered at cell z with fixed observed cases $m =$ (e.g., 1%, 5%, 10%, 20% of total cases in G). Usually, a large m is more sensitive to large clusters and a smaller m is more sensitive to small clusters. Note that the circle Z is an aggregation of several cells, including the center cell z . The expected

number of cases for each circle Z is then $E_z = \frac{c_+}{n_+} n_z$. Assuming $C_Z \sim \text{Poisson}(E_z)$, we obtain the probability of the random variable $C_Z > c_z$. If the probability is smaller than 0.05 (i.e.,

$1 - \sum_s^{c_z-1} \frac{e^{-u(z)} u(z)^s}{s!} > 0.95$), the circle zone Z centered at cell z has excess events and the observed cases in cell z (c_z) will then be added into the statistic R .

3.3 Test statistics for cluster detection methods

Cluster detection methods involve local assessment of the spatial correlation and focus on identifying collections of cases inconsistent with the null hypothesis of no clustering and evaluating their level of significance.

3.3.1 Likelihood based methods—We include four methods (FSS, ULS, SaTScanO, and SaTScanE) using likelihood ratio theory to obtain the statistics. For a given zone Z , the likelihood is defined as

$$L(Z) = \left\{ \left(\frac{c_z}{n_z} \right)^{c_z} \left(\frac{n_z - c_z}{n_z} \right)^{n_z - c_z} \left(\frac{c_{Z'}}{n_{Z'}} \right)^{c_{Z'}} \left(\frac{n_{Z'} - c_{Z'}}{n_{Z'}} \right)^{n_{Z'} - c_{Z'}} \right\} I \left(\frac{c_z}{n_z} > \frac{c_{Z'}}{n_{Z'}} \right),$$

where the Z is the collection of all the possible cells (z 's), with varying shapes and geographic sizes in study region G , and Z' is the complement, i.e., $Z' = G - Z$. ($c_z, c_{Z'}$) are the numbers of cases inside and outside of Z , respectively, and ($n_z, n_{Z'}$) are the corresponding populations. Note that we are only interested in identifying clusters of cells with higher rates, so the indicator function is included. If we wish to study clusters of both high and low rate cells, the indicator function will be removed in $L(Z)$. The statistic for all

four methods is $\lambda = \frac{\text{Max}_{z \in Z} L(z)}{\text{Max}_{z \in G} L(z)}$, or $\log(\lambda)$ as a monotone function of λ , where the denominator is independent of the search zone Z , so that the Z that maximizes the numerator also maximizes the statistic λ . This zone Z is called the maximum likely cluster.

The difference among the four methods is the definition of the search zone Z (scan window). For aggregated count data, a single centroid is defined at each cell z in the SaTScanO method. Around each centroid, we construct one or more areas as a series of hierarchically overlapping areas of monotonically increasing size, such as several different overlapping circles of different sizes, all centered at the same centroid. In SaTScanE, we explore not only circular areas, but also ellipses. The areas covered by one circle with radius may be covered by several ellipses with the long radius δ , but a different short radius and several different angles. Defining the shape of ellipses by the ratio of long radius to short radius, we use the combination of shapes 1 (circle), 1.5, 2, 3, and 4 in our analysis. Once the center of the additional region enters the scan window when we increase the window size, we include this region in the window. The flexible scan statistic (FSS) imposes an irregularly shaped scan window in the study area by connecting its adjacent regions. The method starts at each cell z in the study region and then creates a set of irregularly shaped windows within $K - 1$ nearest neighbors of the starting cell. Thus, for any given cell z , the circular spatial scan statistic considers many concentric circles as the scan area, whereas the flexible scan statistic considers all the concentric circles with connected cells inside plus all the sets of connected regions (including single cell z) with centroids located within the largest concentric circle.

Both the flexibly shaped and elliptic scan methods search more zones than circular scan method, but also require more computing time. The computational load is particularly heavy for the flexibly shaped scan method; according to the authors [26], this method allows a maximum of 30 cells in the cluster region (a limitation of this method compared with others). Therefore, we only use FSS on the state-level data with 49 cells, but not county-level data (3109 cells). When the 50% maximum search window is applied, we have less than 30 cells in the scan window.

Instead of using general criteria to specify the scan windows that are the same for all data sets with the same geographic cells, the upper level set (ULS) scan method for defining the collection of search zones (Z 's) is data dependent. This is a unique feature of this adaptive method compared with other cluster detection methods. The upper level sets are determined by the data from the empirical cell rates ($r_z = c_z/n_z$). For one data set, we first define g as the maximum of the r_z over all z , then the second highest value, third highest value, and so on. The cells (either adjacent or not) with $r_z > g$ are called the islands. We obtain the smallest value for g when the population of all the islands with $r_z > g$ will exceed 50% of the total population if one more cell is added to the collection of the islands at this value of g . For each g , we have a ULS that consists of all islands at this g level. The zones containing connected cells are defined as the candidate zones in ULS. Disconnected zones are treated as different zones in ULS. The zones obtained from all the g 's are the candidate zones (Z 's) for likelihood ratio comparison in one data set. The ULS is not the same for different data sets, because the g depends on the set of cell rates in each dataset.

For all the above four methods, the area (scan window) size could be a fixed population size or a fixed geographic size. The maximum scan window usually is defined to be less than or equal to 50% of the total population or geographical size. If we are interested only in local cluster patterns, we use a smaller maximum scan window. If we are searching for more global spatial patterns, we define a larger maximum scan window.

3.3.2 Turnbull et al's cluster evaluation permutation procedure (CEPP)—Instead of searching for clusters over all possible shapes and sizes with some upper limitations as described in SaTScan, FSS, and ULS, CEPP looks for clusters of a fixed size N^* , which is the user-defined total number of persons at risk. CEPP may detect different clusters under different N^* 's. We create the search window starting from cell z ; then we aggregate cell z

and the cells nearest the centroid of cell z until reaching the total of N^* persons at risk. The search windows vary in geographic size but maintain a constant size of population at risk. We calculate the rates of cases/population for all the windows of the same population size as the area measures. Since the population size is the same, the measure is simplified to be the case counts. The statistic λ is then the maximum of the cases over all the scan windows for CEPP.

3.3.3 Local indicators of spatial association (LISA)—Moran's I statistic, a commonly used measure of overall clustering, can be decomposed to create a standard

normal score for each place z : $\frac{I_z - E(I_z)}{\sqrt{Var(I_z)}}$ where

$$I_z = \left(\frac{1}{s^2} \right) \frac{(Y_z - \bar{Y}) \sum_{z'=1}^T w_{zz'} (Y_{z'} - \bar{Y})}{\sum_{z'=1}^T w_{zz'}}$$

is the contribution from place z to the overall Moran's I statistic. The locations with unusual normal scores will be treated as clusters. This method is one implementation of the class of Local indicators of spatial association (LISA), which can be used for cluster detection. However, we will not use the normal score to assess significance because the Monte Carlo testing procedure will be used for all the methods discussed in this article. We simply use I_z as the statistic for cell z , $z = 1, \dots, T$. We only evaluate the method with Y as the number of cases, and we denote it as LISA regular version (LISAregr). Similar to the variations of the global Moran's I considered here, we will use

$$I_z^* = \left(\frac{Y_z - rn_z}{\sqrt{rn_z}} \right) \frac{1}{\sum_{z'=1}^T w_{jk} z'=1} \sum_{z'=1}^J w_{zz'} \frac{(Y_{z'} - rn_{z'})}{\sqrt{rn_{z'}}},$$

as the local statistic for each cell, where $Y_z = c_z$. We call this the LISA normalized version (LISAnorm). Note that all the cluster detection methods described in 3.3.1 and 3.3.2 are based on evaluating Z over aggregated cells. However, LISA evaluates cells directly.

3.4 Hypothesis testing and significant clusters

Under the constant risk hypothesis, we have $H_0: C_z \sim \text{Poisson}(\tau \times n_z)$, $z = 1, \dots, T$. C_z is the random variable for the aggregated counts in z , and with constant risk τ for all the cells. The relative risks (τ_z^* , $z=1, \dots, T$) for all z 's are then 1 under the constant risk hypothesis (no clustering). Conditioning on a fixed total number of cases, the null hypothesis becomes

$$H_0: C_1, \dots, C_z \sim \text{multinomial}(c_+, n_1/n_+, \dots, n_T/n_+),$$

and the alternative is then

$$H_a: C_1, \dots, C_z \sim \text{multinomial}(c_+, \frac{\tau_1^* n_1}{n_+^*}, \dots, \frac{\tau_T^* n_T}{n_+^*}),$$

where $n_+^* = \sum_{z=1}^T \tau_z^* n_z$ and the relative risk τ_z^* is 1 for the location with the lowest risk of

events (z_0) and larger than $1(\frac{\tau_z}{\tau_{z_0}} > 1, z \neq z_0)$ for locations with a higher risk of events. If the relative risks are not homogeneous, we reject the null hypothesis and claim there is a tendency of clustering and that the locations associated higher relative risks are possible clusters.

For observed data, we compute the p-values for all the tests except Tango's MEET through a Monte Carlo procedure under the constant risk hypothesis. The statistic λ for Tango's MEET itself is a p-value, so no further computation is required. That is, if λ from Tango's MEET is smaller than 0.05, there is a significant global clustering pattern. Note that in this article, we simulate observed data as data with cluster patterns described in section 2. More details of the simulation are included in the next section. For all the methods except Tango's MEET, we simulate multiple count data under H_0 , assuming homogeneous relative risk in G and with fixed total number cases equal to that in the corresponding observed data. These data are null data sets, and the λ 's for the null data sets provide an empirical distribution of λ under the null hypothesis. We reject the null hypothesis if the rank of the observed statistic is larger than $\alpha \times (M+1)$, where M is the total number of simulations with fixed total cases. Usually $\alpha = 0.05$. The p value is

$$p = 1 - \frac{\# \text{ of } (\lambda \text{ observed} > \lambda \text{ simulated})}{1+M}.$$

For LISA, there is a statistic λ_z and corresponding p_z for each cell z . The p_z can be written as

$$p_z = 1 - \frac{\# \text{ of } (\lambda_z \text{ observed} > \lambda_z \text{ simulated})}{1+M}, z \in G.$$

The z 's with p_z smaller than 0.05 are called significant clusters. We did not adjust for multiple testing in our study. For all the other cluster detection methods, the areas covered by the scan window (Z) associated with p-values smaller than 0.05 are significant clusters. We can also order the values of the $L(Z)$ or the cases and obtain not only the maximum value of λ , but also the ones smaller than the maximum value as $\lambda^2, \dots, \lambda^L$. If $\lambda^2, \dots, \lambda^L$ from the observed data are also larger than the 95th percentile of the λ from the null data sets, the Z 's associated with the $\lambda^2, \dots, \lambda^L$ are called secondary clusters, significant at the 0.05 level.

3.5 Simulation, power, and precision

To evaluate the performance of the spatial tests discussed in section 3.2 and section 3.3 on detecting clustering and the locations of clusters for population cancer cases data, we simulated 1000 county-level alternative data sets (observed data) under H_a with a fixed number of total cases, assuming heterogeneous relative risks in the United States; and 10,000 county-level null data sets under H_0 with the same fixed total number of cases, assuming constant relative risk. The real county-level population in 2000 was used in the simulation. We tested the eight spatial patterns (sites) described in section 2. For each pattern, we allowed the c to be 0.1, 0.2, 0.5, and 1.0, so the maximum relative risk for the cluster regions will be 1.1, 1.2, 1.5, and 2.0 respectively. We also varied the sample size (2500, 5000, 10000, 25000, and 50000) to represent a range from rare to common cancers. For very rare cancers with only a few hundred cases observed across the whole country, data are often aggregated over several years to reduce the variability of the number of cases. It is usually possible to aggregate a sufficient number of years to total 2500 cases even for very

rare cancers. Under H_a , there are 160 combinations of dataset characteristics (8 sites $\times$ 4 relative risk levels $\times$ 5 sample sizes), and under H_0 , there are 40 combinations (8 sites $\times$ 5 sample sizes). For each combination, we compute 1000 λ 's for the alternative data and compare each with the 10000 λ 's for the corresponding null data. We obtain one p-value for each λ from the 1000 (observed) alternative data sets. Then at 0.05 level, the power of rejecting the null hypothesis of no clustering (constant risk) is (# of p-values < 0.05) / 1000 for each of the 160 combinations under H_a . Note that for Tango's MEET, the power is directly computed as (# of λ 's < 0.05) / 1000.

The data are simulated at the county level. To implement the methods on state-level data, we simply aggregate the simulated county data to the state level. We illustrate the spatial pattern of cervical and colorectal cancer sites at the state level in Figure 2. The relative risks at state level in the two maps are the weighted averages of the relative risks at county level, with county population as the weight.

In the data simulated from the multinomial distribution, we can not control the overdispersion parameter. However, the clustered structure of the underlying cancer patterns used as the basis for the simulations induced overdispersion. The estimated overdispersion parameter (deviance / DF) as discussed in [37] varies for different sites, sizes and relative risks. We calculate the overdispersion parameter for all the 1000 simulated data under each strata (site/relative risk/sample size), the average value is obtained over the 1000 simulations for each strata. For example, for colorectal cancer with a relative risk of 2 inside clusters, the average overdispersion parameter is 0.77, 0.97, 1.18, 1.54 and 2.04 for sample size 2.5k, 5k, 10k, 25k and 50k, respectively; for sample size 50k, the average overdispersion parameter is 1.07, 1.11, 1.34, and 2.04 for relative risk 1.1, 1.2, 1.5 and 2.0, respectively. Some of the estimates are smaller than 1, which indicates underdispersion due to the presence of many simulated 0 observations for the smaller populations. The overdispersion estimate also varies by sites, for example, for liver cancer with sample size 50k, the average overdispersion parameter is 1.08, 1.13, 1.48, and 2.43 for relative risk 1.1, 1.2, 1.5 and 2.0, respectively. The values for liver are slightly higher than the ones for colorectal site.

When any clusters are detected, we also evaluate the precision of those clusters, i.e., that the tests detected clusters in the correct locations. Precision is measured using either the ratio of the number of cells detected in significant clusters that are in the true clusters to the total number of cells in the true clusters (r_T) or to the number of cells in detected clusters (r_D). If no significant clusters are detected, the overlapping area is zero, yielding ratios (r_T and r_D) of zero. Thus, the values of r_T and r_D are between 0 and 1, with 0 indicating poor precision and 1 perfect precision. These measures are analogous to sensitivity and positive predictive value (PPV), respectively, e.g., a large value of r_T implies the method finds most of the true cluster locations, and a large r_D indicates fewer false positive cells (cells outside true clusters that are claimed to be in clusters). Both measures are important indicators of the precision of the cluster detection method. For example, a method that declares the entire geographic area to be a significant cluster will have high sensitivity (the smaller true cluster will be included) but low PPV (most of the detected cluster is not part of the true cluster), making it less useful than a method that is high by both measures. For each of the 160 combinations, we can compute the average of r_T ($\bar{r}_T$), the average of r_D ($\bar{r}_D$), the standard deviation of r_T and r_D over the 1000 repeats, and the corresponding 95% confidence intervals as $(\bar{r}_T - 1.96 \times \text{std}(r_T), \bar{r}_T + 1.96 \times \text{std}(r_T))$ and $(\bar{r}_D - 1.96 \times \text{std}(r_D), \bar{r}_D + 1.96 \times \text{std}(r_D))$ for each combination.

The computation for SaTScanO and SaTScanE are completed by using the SaTScan software (www.satscan.org). It is very hard to retrieve the location information of the detected clusters for all the iterations using the current software, so we only compute the

precision measures r_T and r_D for a randomly selected data set from the 1000 data sets in each combination. All the other statistics are computed by our own program written in C++. We can compare the point estimates of r_T and r_D from SaTScan methods with the confidence intervals of the measures from other methods to rank the performance of the methods by precision of the detected clusters. If the r_T and r_D from either SaTScanO or SaTScanE are larger than the $\bar{r}_T$ and $\bar{r}_D$ from the other methods, there is an advantage in terms of precision for the SaTScan methods, particularly if the r_T and r_D from SaTScan are even larger than the upper level of the 95% confidence intervals from the other methods. For the comparison among all the other methods except SaTScan, we can directly compare the $\bar{r}_T$ and $\bar{r}_D$, and also evaluate the variability of the methods by comparing the width of their confidence intervals.

We also developed maps to present the location of detected clusters from multiple replicates (1000). For all of the discussed cluster detection methods except SaTScanO and SaTScanE, we assign 1 if the cell is in the detected cluster areas and 0 outside for each simulated alternative data set, then we sum the values over the 1000 repeats. The summed values are between 0 and 1000 for each cell, with a higher sum indicating more frequent detection of cells as a part of the detected clusters. For example, if a cell was found to be inside the detected clusters for all 1000 replicates, it has 100% probability of being in the detected cluster areas. Similarly, if a cell has never been included in a detected cluster, the value is zero for this cell. The varying levels of the probability of inclusion of each cell in a detected cluster is represented by 6 levels of color, from light to dark (with sums categorized as 0, 1–200, 201–400, 401–600, 601–800, and 801–1000). The areas with darkest shading can be treated as the most frequent locations estimated for the cluster areas; the other lighter colored areas reflect the variability of the methods in detecting the cluster, i.e., a type of spatial confidence interval. For SaTScan methods, only one random data set from the 1000 repeats is evaluated, so there are two colors (light for 0 and dark for 1) in their maps, 0 implies outside the detected clusters and 1 otherwise. The p-values for detected significant clusters from SaTScan are also presented, which reflect the probability of inclusion of the detected cells in the true cluster areas. The precision of the cluster detection methods can be judged visually by comparing these replicate cluster maps to the original (true) patterns in Figure 1 and 2.

4 Results

4.1 Comparison of powers for methods with several available versions

Moran's I—Three versions of Moran's I were included in this study (normalized version, rate version, and regular version). Moran's I normalized version is the best of these versions in terms of power, followed by Moran's I regular version, as shown in Table 2. Moran's I rate version has very poor power for all combinations. Power for Moran's I regular version is fair for two patterns where high rates are clustered in high-population areas (bladder and liver have the highest average population in all cluster areas, as shown in Table 1). Only when the spatial pattern of cancer cases is consistent with the population spatial pattern can the regular version provide reasonable power, and then only for clusters in high population areas. Large numbers of observed cases are usually due to a large local population or a high relative risk. The regular version of Moran's I has lower power to detect clustering when the relative risk is high but the local population is low, e.g., for the male lung pattern. It may also produce many false-positive locations due to the variation in population instead of varying relative risk of the disease. The rate version improves on the regular version by dividing the number of cases by the population, but low-population areas usually have the most variable rates, so that variability of the rates depends on both population size and spatial structure in cancer cases. This rate version is not appropriate for use for cancer data

from heterogeneous populations. The normalized version tests whether residuals from expected cases based on constant risk hypothesis are spatially correlated. Because this overcomes the problem of variability in population and rates, the normalized version has better power than the other two versions.

Tango's statistics—We evaluate Tango's EET based on adjacent weights (TangoADJ) and Tango's MEET based on population density exponential weights (TangoPDM). The powers for TangoPDM are higher than or equal to those for TangoADJ for all combinations. The advantage of TangoPDM is the way it formulates the weight function. In TangoADJ, the weight is 1 if two cells are adjacent and is 0 otherwise, which focus on evaluating the spatial correlation among neighbors. However, the TangoPDM in our study allows the weights to change with the parameter ranging from 5% to 50% of the total population, and the optimal p-value is obtained over all the selection of weights and recorded as MEET. In this way, the TangoPDM method evaluates the clustering tendency in areas with varying sizes. The construction of the statistics indicates that better power will be achieved by TangoPDM compared to TangoADJ.

Besag-Newell's R—In Besag-Newell's R, the user must define the size of the circle centered at each cell z . We used 5%, 10%, and 20% of the total number cases in our study (BN05, BN10, and BN20). For the bladder, colorectal, and lung (female) cancer sites, the statistics with 5% window size provide the best power of the three sizes. As shown in Table 1, the bladder, colorectal, and lung (female) sites have smaller numbers of counties in their true clusters (356, 378, and 324, respectively) than the other sites. For the other five sites, the statistics with 10% window size provide the best power of the three sizes. The optimal window sizes (5% and 10%) are smaller than all the proportions of expected cases for all the corresponding sites (Table 1, last column), because the statistic in this method is the sum of the observed cases for the cells that each has unusually high observed cases around its neighborhoods with varying sizes (5%, 10%, and 20%). We do not add the total number of observed cases in the search window centered at a cell into the statistic, but only the observed cases in the particular centroid cell. A search window with larger radius will have better power than a window with a smaller radius with the same centroid, if the whole window is inside the true cluster region. However, a larger window may also have a greater chance of including more cells outside the true cluster areas; in this situation, we lose the power to add the observed cases of a cell as the center of the window into the statistic R, even when the centroid cell is located in the true cluster region. Therefore, large window size is not preferred in this method. We also observed that although optimal power was achieved at 10% in Table 2, the power at 5% is still not bad and close to those at 10%. When the optimal power is realized at 5%, some powers at 10% or 20% are much lower than those at 5%. Therefore we suggest a small window size when using Besag-Newell's method in practice.

CEPP statistics—The CEPP method also has a user-defined parameter. The search window (circle) is defined to have 5%, 10%, and 20% of the total population in our study (CEPP05, CEPP10, and CEPP20). It turns out that the optimal window size is associated with the proportion of the population in the cluster areas of the total population. CEPP20 has the best power for the cervix, kidney, and lung (male) sites (Table 2), all three sites have proportions of population in the clusters larger than 10% (14% for cervix; 19% for kidney and for lung male), and all have large number of counties in cluster areas. CEPP10 has the best powers for the bladder, colorectal, and prostate sites (the proportion of population is 16%, 10%, and 10%, respectively). Note that it is easier to cover the clusters in the cervix site with a large circle compared with bladder, so a 20% window is preferred for cervix, whereas a 10% window is preferred for bladder. CEPP05 has better powers for liver and

lung (female) (with population proportions at 16% and 9%, respectively). The powers for 5%, 10%, and 20% at the liver site are very close, even though the best is 5%. Because it is hard, in practice, to decide the optimal size in this method, researchers usually evaluate CEPP with several window sizes.

4.2 Comparison of powers across the tests

Based on the results in Table 2, we classify the methods into four groups in terms of the average power by site, from high to low. The best power for all sites is SaTScanE (average powers >80%). Group No. 2 includes TangoPDM (average powers >70%). Group No. 3 includes SaTScanO, CEPP, Besag-Newell's method, and FSS, with the average powers greater than 60%. Group No. 4, with the worst power, includes TangoADJ, ULS, and Moran's I normalized version (average powers >40%). A similar rank ordering is observed for the tests when the average power is stratified by sample size and by relative risk in this table. Note that all the methods except FSS are applied to county-level data.

The performance of CEPP and Besag-Newell's method depends on the user-defined parameter (size of the scan window). We use the best powers out of the three selected sizes, which artificially improves the power of the two methods. In practice, these methods may not always be better than the other methods in the same group.

The performance of ULS is poor for county-level data. A possible reason is that there are 3109 counties in the study region, and the population is very heterogeneous. Therefore, with a fixed total number of cases, many counties will not have cases observed. There is no study of the performance of ULS on data with large cells and sparse data. From our results, it appears that only adjacent cells are considered to be neighbors, ULS may have trouble connecting the possible cells to form large islands above the level in the true cluster areas when we have sparse data with many empty cells. When we aggregate the county-level data to state-level data and apply the ULS method, the power of ULS improves and approximates the power of the methods in Group No. 3 (SaTScanO, CEPP, Besag-Newell). This improvement is probably due to the fact that it is nearly impossible to have empty cells with a total number of cases of 2500 and above when there are only 49 cells (states). ULS seems to require large sample sizes for data with large geographic units (cells), especially when the population is not homogeneously distributed over space. With limited sample size, it is always better to work with ULS on larger geographic units. FSS also uses adjacency in neighbor definition, so may have the same pattern as ULS has on sparse data.

4.3 Powers for the varying sites/patterns

Test performance is associated with four factors: number of geographic cells included, population density (or average population), urban-rural status, and concentration of the multiple clusters. The shape of a single cluster is not as important as these four factors, because we can always use a large window to cover several small clusters in the cluster detection methods, and global clustering methods are not affected by the shapes. Tests applied to the bladder pattern have the best power, because its clusters are concentrated in large population areas of the northeast (average population 127,000), covering more urban areas, even though there is only 356 counties in the cluster areas, as shown in Table 1. The power for the cervix, kidney, liver, and lung (male) sites are close overall. The prostate site has lower power than the above sites, but better power than colorectal and lung (female). The powers for colorectal and lung (female) sites are very poor compared with the others because of their small county numbers, low population and scattered multiple clusters. From Table 2, a site includes clusters with either large population size or large geographic size (more counties), covers more urban areas, and has other clusters nearby, the power of detecting the clusters is higher. Edges of the clusters do not seem to affect the power.

4.4 Powers for varying sample sizes and relative risks

As shown in Table 2, the average powers by relative risk (RR) and sample size (SS) increase with sample size and relative risk in cluster areas for all methods. These patterns can also be observed in the selected plots (Figure 3). The powers by site, relative risk and sample size (160 combinations) can not be included in the text because of the space limitation and are available in the appendix. Note that the performance of ULS and Moran's I rate and regular versions have very unstable performance and poor power on the county-level data. For methods in Group No. 1 (SaTScanE), good power (90%–100%) is achieved at the combination of (RR=1.1 and SS=25,000) for two sites with the highest average population in the whole cluster areas (bladder and liver), and (RR=1.1 and SS=50,000) for other sites. SaTScanE also has good powers for data with RR as 1.2, the average power reaches 93.7 at (RR=1.2 and SS =10,000). When the relative risk is 1.5 or 2.0, the powers for all sites and sample sizes are good for SaTScanE.

For methods in Group No. 2 and 3, good average power (90%–100%) is achieved at the combination of (RR as 1.2 and SS as 25,000+) or (RR as 1.5 and SS as 5,000 +) or (RR as 2.0 and all sample sizes). For cases with relative risk as 1.1, only for sites bladder, cervix, liver, lung male and kidney, those methods can have power larger than 90% at sample size 50,000.

The average powers for the Moran's I normalized version and TangoADJ (Group 4) have good powers at the combination of (RR=1.2 and SS=50,000) or (RR=1.5 and SS=10,000+). For ULS, the average power reaches 90.7 at (RR=1.5 and SS=25,000). The power for Moran's I rate and regular versions is poor overall.

4.5 Precision of detected clusters

In the previous section, we notice that the powers for methods in Group 1, 2, and 3 is as high as 90%–100% at moderate sample size such as 10,000–25,000, and the power is close to 100% at larger sample sizes for a relative risk level of 1.2. Here, we evaluate the precision for data with relative risk 1.2 and sample size 50,000 in order to investigate the chance of the detected cluster covering the true cluster areas when the power is good. We also provide the precision measures at 5000 when the power is poor for comparison. Table 3 presents the r_T and r_D , hereafter referred to as sensitivity and PPV, respectively, and their corresponding confidence intervals over the eight sites and for 5000 and 50,000 cases.

At 5k, the sensitivity and PPV for a randomly selected alternative data set for SaTScan methods are all zero (no significant cluster detected), except that a small portion of the true cluster was detected for the prostate site. The intervals for the other methods are wide and the lower limits are all zero except LISA. However, the upper limits of the confidence intervals for LISA are all very low (lower than 0.3). Therefore, the precision is poor for all methods at 5000 sample size for data with relative risk 1.2, which is consistent with the power performance.

At 50,000, SaTScanO has better sensitivity than SaTScanE, but lower PPV. For methods where intervals could be computed, CEPP20 has the highest sensitivity (more true cluster areas are found). The LISA regular version has the smallest variation, but its sensitivity is very low. The variation of the statistics from the ULS method on county-level data is very large (the length of the CI is about 0.70). The sensitivity measure for SaTScanE is within the interval of that for CEPP20, but with much larger PPV, i.e., fewer false positives. The SaTScan methods have the best precision for county-level data.

For the CEPP methods, the sensitivity increases when the window size of CEPP increases from 5% to 20% of the total population, because we have a better chance of including the

true cluster in the detected cluster areas with a bigger window. Nevertheless, the r_D increases when the window size increases, implying that more false positives occur when a bigger search window is used. The high PPV for CEPP05 does not indicate the method is good, because its sensitivity is very small, which implies that the detected cluster areas are only a small piece of the true cluster areas even though the false-positive rate is low.

At the state level, the precision for ULS and FSS is very good, especially because both have very small false-positive rates (large r_D). FSS has lower sensitivity, but higher PPV and smaller variation across data sets compared with ULS. We can not compare the precision of the county-level and state-level results because the geographic units are not the same.

We selected two sites to produce the maps for illustration (cervix from a group with good power and colorectal from a group with poor power). We also evaluate the ULS method on county-level data with sample size 300,000 here, because it is close to 100 times the number of counties. For state-level data, we use 50,000 as sample size, which is also close to 100 times the number of states. As seen in the maps (Figure 4 and 5), ULS has good precision for county-level data with sample size of 300,000 and for state-level data with sample size 50,000 as well. In terms of precision, if the sample size is large enough, both ULS and FSS have good performance in detecting the cluster locations with very few false positives. For methods with compact search windows (e.g., SaTScan methods and CEPP), the chance of detecting clusters is good (high sensitivity), but so is the chance of finding many false positives (PPV may be small). LISA (normalized) only detected tiny spots of the true cluster locations with low power, but the false-positive rates are low. LISA (regular) has a better chance of detecting pieces of the true cluster areas, but it also detects some areas far away from the true cluster region, as shown in both the cervix and colorectal maps.

Overall, in terms of precision, ULS and FSS are better if the sample size is very large and the total number of cells in the cluster is small. However, the performance of ULS is not good for sparse data, and the FSS cannot work for county-level data because of the limitation in computation speed. Therefore, SaTScan methods become a suitable choice with a relatively good chance of finding most of the true cluster areas (high sensitivity) and lower false-positive rates (high PPV).

5 Discussion

Two types of tests are evaluated in this paper (global clustering and cluster detection), which are designed for studying spatial heterogeneity. Global clustering methods test overall global spatial correlation, and cluster detection methods identify unusual collections of cells compared with others. A significant local cluster does not always imply significant global clustering tendency, and vice versa. In this study, we simulate data with clusters by simulating unusual collections of cells with high relative risks compared with others. Therefore, cluster detection should work well for the data simulated in this paper. In addition, because all sites have multiple clusters that cover many places, our clusters should be large enough for the global methods to declare significant global clustering in all eight patterns. In terms of power, the best method is TangoPDM among the tests for global clustering and SaTScanE for cluster detection methods. The power for SaTScanE is slightly better than TangoPDM for the simulated alternative data with the eight patterns. The minimum sample size to achieve good powers (>90%) varies with relative risk, tests, and cluster patterns. We will have good power for data with sample size 10,000 for SaTScanE and 25,000 for TangoMEET, when the relative risk is at 1.2 level.

Note that Moran's I has the worst performance among all the methods tested. Although Moran's I statistic was proposed over 50 years ago, its statistical properties have only

recently been examined. Walter [31] found that the power of Moran's I statistic based on rates was slightly better than a similar statistic, Geary's c, and demonstrated that population heterogeneity must be taken into account in order to properly identify disease rate clusters. Assuncao and Reis [19] proposed a standardized version of Moran's I and found that it had better power than the original test, which had low power unless populations were held constant. Kulldorff's recent papers also showed that Moran's I had low power [12,38].

Except for the proposed modification by Assuncao and Reis, all of these studies to date have compared power for the rate version of Moran's I, purportedly to account for population heterogeneity by dividing the number of cases by the corresponding population. However, as Assuncao and Reis (1999) ([23]) point out, rates based on small populations have greater variability and thus are more likely to be extremely high values. The standardized version of Moran's I adjusts for these population variances. We found that this version had better power than the original form, based on either counts or rates, and was relatively robust to differences in pattern. However, even with the best version of Moran's I (normalized version), the power is still poor compared with other methods.

Advantages of our work are that realistic cluster patterns of cancer are studied, and, since we use simulated data, the true clusters are known so that methods can be compared. However, the patterns studied in this article are limited to observed large spatial clustering patterns from cancer data, not very small clusters, outliers (anomalies), or clusters with very elongated shapes, e.g., connected irregular lines as along a river. It is possible that the performance is different for other spatial patterns that are not evaluated in this paper. In addition, the levels of under- and over-dispersion in our simulated data may not reflect a full range of levels seen in actual data.

Geographic units (cells) affect the results (power and precision). The focus of our study is not on the effect of the definition of geographic cells on the power to detect global clustering, even though we notice that the performance of tests on varying definition of cells is different. The methods that do not depend on the adjacency matrix may have bigger power when there are more cells (from state to county), but methods involving neighbors may have lower powers when the smaller cells are used (county). Because there are many empty cells when the cell number is large with a limited total number of cases. It may be interesting to study performance of the test on data with varying geographic units.

Acknowledgments

We thank Mark Hachey, Jun Luo and Jeremy Lyman from Information Management System, Inc. for technical assistance. We also thank Dave Stinchcomb from NCI for helping in adjacency matrix generation.

Reference

1. Pollack LA, Gotway CA, Bates JH, Parikh-Patel A, Richards TB, Seeff LC, Hodges H, Kassim S. Use of the spatial scan statistic to identify geographic variations in late stage colorectal cancer in California (United States). *Cancer Causes and Control*. 2006; 17:449–457. [PubMed: 16596297]
2. Fukuda Y, Umezaki M, Nakamura K, Takano T. Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan. *International Journal of Health Geographics*. 2005; 4:16. [PubMed: 15953394]
3. Han DW, Rogerson PA, Nie J, Bonner MR, Vena JE, Vito D, Muti P, Trevisan M, Edge SB, Freudenheim JL. Geographic clustering of residence in early life and subsequent risk of breast cancer (United States). *Cancer Causes and Control*. 2004; 15:921–929. [PubMed: 15577294]
4. Buntinx F, Geys H, Lousbergh D, Broeders G, Cloes E, Dhollander D, Op De Beeck L, Vanden Brande J, Van Waes A, Molenberghs G. Geographical differences in cancer incidence in the

- Belgian province of Limburg. *European Journal of Cancer*. 2003; 39:2058–2072. [PubMed: 12957461]
5. Waller, LA.; Gotway, CA. *Applied Spatial Statistics for Public Health Data*. New Jersey: John Wiley and Sons, Inc; 2004.
 6. Knorr-Held L, Rasser G. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*. 2000; 56(1):13–21. [PubMed: 10783772]
 7. Besag J, York J, Mollie A. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*. 1991; 43:1–59.
 8. Richardson S, Thomson A, Best N, Elliot P. Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives*. 2004; 112:1016–1025. [PubMed: 15198922]
 9. Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. *Stat Methods Med Res*. 2005; 14(1):35–59. [PubMed: 15690999]
 10. Waller LA, Hill EG, Rudd RA. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. *Statistics in Medicine*. 2006; 25(5):853–865. [PubMed: 16453372]
 11. Kulldorff M, Song C, Gregorio D, Samociuk H, DeChello L. Cancer map patterns: are they random or not? *American Journal of Preventive Medicine*. 2006; 30(2):S37–S49. [PubMed: 16458789]
 12. Song C, Kulldorff M. Power evaluation of disease clustering tests. *International Journal of Health Geographics*. 2003; 2(1):9.
 13. Kulldorff M. Tests of spatial randomness adjusted for an inhomogeneity: a general framework. *Journal of the American Statistician Association*. 2006; 101(475):1289–1305.
 14. Moran PAP. Notes on continuous stochastic phenomena. *Biometrika*. 1950; 37:17–23. [PubMed: 15420245]
 15. Besag J, Newell J. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*. 1991; 154:327–333.
 16. Tango T. A class of tests for detecting ‘general’ and ‘focused’ clustering of rare diseases. *Statistics in Medicine*. 1995; 14:2323–2334. [PubMed: 8711272]
 17. Tango T. A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine*. 2000; 19:191–204. [PubMed: 10641024]
 18. Geary RC. The contiguity ratio and statistical mapping. *The Incorporated Statistician*. 1954; 5:115–145.
 19. Oden N. Adjusting Moran’s I for population density. *Statistics in Medicine*. 1995; 14:17–26. [PubMed: 7701154]
 20. Cuzicki J, Edwards R. Spatial clustering for inhomogeneous populations (with discussion). *Journal of the Royal Statistical Society, Series B*. 1990; 52(1):73–104.
 21. Swartz JB. An entropy-based algorithm for detecting clusters of cases and controls and its comparison with a method using nearest neighbors. *Health and Place*. 1998; 44:67–77. [PubMed: 10671012]
 22. Whittemore AS, Friend N, Brown BW, Holly EA. A test to detect clusters of disease. *Biometrika*. 1987; 74:631–635.
 23. Assuncao RM, Reis EA. A new proposal to adjust Moran’s I for population density. *Statistics in Medicine*. 1999; 18:2147–2162. [PubMed: 10441770]
 24. Kulldorff M. A spatial scan statistic. *Communications in Statistics: Theory and Methods*. 1997; 26:1481–1496.
 25. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine*. 2006; 25(22):3929–3943. [PubMed: 16435334]
 26. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*. 2005; 4:11. [PubMed: 15904524]
 27. Turnbull GW, Iwano EJ, Burnett WS, Howe HL, Clark LC. Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. *American Journal of Epidemiology*. 1990; 132:136–143. [PubMed: 2356805]

28. Patil GP, Taillie C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*. 2004; 11:183–197.
29. Anselin L. Local indicator of spatial association. *Geographical Analysis*. 1995; 27:2.
30. Duczmal L, Assuncao R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis*. 2005; 45:269–286.
31. Kelsall JE, Diggle PJ. Kernel estimation of relative risk. *Bernoulli*. 1995; 1(1/2):003–016.
32. Duczmal L, Kulldorff M, Huang L. Evaluation of spatial scan statistics for irregular shaped clusters. *Journal of Computational and Graphical Statistics*. 2005; 15(2):1–15.
33. Population Reference Bureau. U.S. counties by urban-rural continuum codes in 2003. available from <http://www.prb.org/rfdcenter/USACountiesBealeCodes2003.pdf> (Most recent visit date is March 25, 2008)
34. Waldhör T. The spatial autocorrelation coefficient Moran's I under heteroscedasticity. *Statistics in Medicine*. 1996; 15(7–9):887–892. [PubMed: 8861157]
35. Walter SD. The analysis of regional patterns in health data. II. The power to detect environmental effects. *American Journal of Epidemiology*. 1992; 136(6):742–759. [PubMed: 1442740]
36. Song C, Kulldorff M. Tango's maximized excess events test with different weights. *International Journal of Health Geographics*. 2005; 4:32.
37. McCullagh, P.; Nelder, JA. *Generalized Linear Models*. New York: Chapman & Hall; 1989.
38. Kulldorff M, Tango T, Park P. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*. 2003; 42(4):665–684.

Appendix

Table 4

Powers (%) of tests on varying cluster patterns (sites), total number of cases (size) and relative risk (Risk).

County level data																	
Site	size	Risk	Besag Newell				Moran's I			Tango		CEPP			Satscan		U
			5%	10%	20%	normalized	rate	regular	ADJ	PDM	5%	10%	20%	circle	ellipse		
bladder	2.50	1.10	11	11	8	6	5	6	6	26	9	9	7	9	32		
bladder	2.50	1.20	46	36	28	9	5	7	16	61	32	37	27	26	63		
bladder	2.50	1.50	100	100	98	48	6	11	89	100	100	100	97	100	100		
bladder	2.50	2.00	100	100	100	100	6	31	100	100	100	100	100	100	100	10	
bladder	5.00	1.10	17	15	14	8	6	8	11	39	19	20	14	12	42		
bladder	5.00	1.20	81	69	52	14	5	6	36	87	71	74	58	64	91		
bladder	5.00	1.50	100	100	100	93	5	17	100	100	100	100	100	100	100	10	
bladder	5.00	2.00	100	100	100	100	6	55	100	100	100	100	100	100	100	10	
bladder	10.00	1.10	39	28	22	8	6	7	15	63	31	38	28	27	65		
bladder	10.00	1.20	100	95	85	35	6	11	71	99	98	98	91	97	100		
bladder	10.00	1.50	100	100	100	100	5	27	100	100	100	100	100	100	100	10	
bladder	10.00	2.00	100	100	100	100	8	79	100	100	100	100	100	100	100	10	
bladder	25.00	1.10	88	75	64	18	5	10	48	95	84	86	68	79	96		
bladder	25.00	1.20	100	100	100	84	5	15	100	100	100	100	100	100	100		
bladder	25.00	1.50	100	100	100	100	5	54	100	100	100	100	100	100	100	10	
bladder	25.00	2.00	100	100	100	100	12	99	100	100	100	100	100	100	100	10	
bladder	50.00	1.10	100	99	95	42	7	10	88	100	100	100	97	100	100		
bladder	50.00	1.20	100	100	100	100	5	20	100	100	100	100	100	100	100		

County level data																
Site	size	Risk	Besag Newell			normalized	rate	Moran's I		Tango		CEPP			Satscan	
			5%	10%	20%			regular	ADJ	PDM	5%	10%	20%	circle	ellipse	UI
bladder	50.00	1.50	100	100	100	100	8	80	100	100	100	100	100	100	100	
bladder	50.00	2.00	100	100	100	100	16	100	100	100	100	100	100	100	100	10
cervix	2.50	1.10	9	11	7	7	6	4	5	20	8	10	11	9	35	
cervix	2.50	1.20	24	26	19	6	5	2	10	41	22	22	27	22	55	
cervix	2.50	1.50	99	96	91	31	4	0	61	99	95	94	98	96	100	
cervix	2.50	2.00	100	100	100	97	8	0	100	100	100	100	100	100	100	8
cervix	5.00	1.10	13	19	14	7	5	3	7	27	16	18	20	13	43	
cervix	5.00	1.20	50	63	48	11	4	2	21	70	48	53	63	41	82	
cervix	5.00	1.50	100	100	100	72	7	0	96	100	100	100	100	100	100	3
cervix	5.00	2.00	100	100	100	100	9	0	100	100	100	100	100	100	100	10
cervix	10.00	1.10	22	37	21	8	5	3	10	44	23	27	35	22	58	
cervix	10.00	1.20	88	92	77	23	6	1	42	95	85	86	91	84	98	3
cervix	10.00	1.50	100	100	100	99	9	0	100	100	100	100	100	100	100	9
cervix	10.00	2.00	100	100	100	100	12	0	100	100	100	100	100	100	100	10
cervix	25.00	1.10	60	77	58	14	5	2	25	82	62	63	75	55	88	3
cervix	25.00	1.20	100	100	99	65	7	1	94	100	100	100	100	100	100	3
cervix	25.00	1.50	100	100	100	100	7	0	100	100	100	100	100	100	100	10
cervix	25.00	2.00	100	100	100	100	18	0	100	100	100	100	100	100	100	10
cervix	50.00	1.10	93	96	90	29	6	1	57	97	93	94	96	93	100	
cervix	50.00	1.20	100	100	100	99	6	0	100	100	100	100	100	100	100	8
cervix	50.00	1.50	100	100	100	100	10	0	100	100	100	100	100	100	100	10
cervix	50.00	2.00	100	100	100	100	37	0	100	100	100	100	100	100	100	10
colorectal	2.50	1.10	7	6	6	6	6	6	6	16	7	6	6	6	29	
colorectal	2.50	1.20	12	11	11	8	5	6	8	25	12	11	13	13	42	
colorectal	2.50	1.50	61	49	35	25	5	5	31	67	62	63	58	96	100	8
colorectal	2.50	2.00	100	99	94	91	6	6	99	100	100	100	100	100	100	9
colorectal	5.00	1.10	8	6	7	7	5	5	6	19	9	11	9	8	35	
colorectal	5.00	1.20	15	15	14	11	4	6	13	29	22	26	22	19	60	3
colorectal	5.00	1.50	93	88	68	58	6	7	75	94	96	96	89	99	100	
colorectal	5.00	2.00	100	100	100	100	7	9	100	100	100	100	100	100	100	10
colorectal	10.00	1.10	11	7	8	7	5	6	7	22	12	13	13	13	47	
colorectal	10.00	1.20	41	26	21	16	6	6	19	50	41	43	38	51	86	
colorectal	10.00	1.50	100	99	96	95	6	10	99	100	100	100	100	100	100	3
colorectal	10.00	2.00	100	100	100	100	8	13	100	100	100	100	100	100	100	10
colorectal	25.00	1.10	25	15	18	9	4	5	14	37	27	29	27	26	68	
colorectal	25.00	1.20	90	70	60	47	5	7	64	88	89	92	83	96	100	3
colorectal	25.00	1.50	100	100	100	100	7	11	100	100	100	100	100	100	100	9
colorectal	25.00	2.00	100	100	100	100	11	23	100	100	100	100	100	100	100	10
colorectal	50.00	1.10	54	38	30	19	6	7	29	61	57	57	51	64	92	3
colorectal	50.00	1.20	100	99	91	90	6	10	98	100	100	100	99	100	100	3

County level data																
Site	size	Rrisk	Besag Newell			normalized	rate	Moran's I		Tango		CEPP			Satscan	
			5%	10%	20%			regular	ADJ	PDM	5%	10%	20%	circle	ellipse	UI
colorectal	50.00	1.50	100	100	100	100	7	16	100	100	100	100	100	100	100	
colorectal	50.00	2.00	100	100	100	100	15	36	100	100	100	100	100	100	100	10
kidney	2.50	1.10	13	14	11	5	4	3	8	24	8	12	12	7	33	
kidney	2.50	1.20	31	40	33	8	5	2	16	55	24	32	37	17	50	
kidney	2.50	1.50	99	99	99	33	6	1	87	100	96	98	99	92	99	3
kidney	2.50	2.00	100	100	100	97	7	0	100	100	100	100	100	100	100	9
kidney	5.00	1.10	14	22	20	8	5	3	10	36	15	17	24	11	43	
kidney	5.00	1.20	54	72	63	12	5	2	32	81	47	63	67	37	70	
kidney	5.00	1.50	100	100	100	70	6	0	99	100	100	100	100	100	100	3
kidney	5.00	2.00	100	100	100	100	8	0	100	100	100	100	100	100	100	10
kidney	10.00	1.10	28	40	31	7	5	2	16	55	23	32	38	21	55	
kidney	10.00	1.20	91	96	91	22	5	0	68	98	84	92	94	77	96	
kidney	10.00	1.50	100	100	100	99	6	0	100	100	100	100	100	100	100	1
kidney	10.00	2.00	100	100	100	100	12	0	100	100	100	100	100	100	100	10
kidney	25.00	1.10	73	84	72	12	4	1	43	92	62	74	80	52	83	
kidney	25.00	1.20	100	100	100	64	6	0	100	100	100	100	100	100	100	4
kidney	25.00	1.50	100	100	100	100	6	0	100	100	100	100	100	100	100	10
kidney	25.00	2.00	100	100	100	100	16	0	100	100	100	100	100	100	100	10
kidney	50.00	1.10	97	100	97	29	6	0	81	100	96	98	98	88	98	
kidney	50.00	1.20	100	100	100	98	5	0	100	100	100	100	100	100	100	1
kidney	50.00	1.50	100	100	100	100	12	0	100	100	100	100	100	100	100	10
kidney	50.00	2.00	100	100	100	100	32	0	100	100	100	100	100	100	100	10
liver	2.50	1.10	8	10	6	5	5	7	7	24	8	9	9	7	29	
liver	2.50	1.20	20	26	12	8	3	9	14	49	23	22	19	15	52	
liver	2.50	1.50	98	98	79	33	4	18	87	100	99	97	93	94	100	
liver	2.50	2.00	100	100	100	98	7	30	100	100	100	100	100	100	100	9
liver	5.00	1.10	11	17	9	7	4	6	10	32	15	14	13	10	38	
liver	5.00	1.20	47	53	28	10	4	12	31	77	52	48	44	38	79	
liver	5.00	1.50	100	100	98	75	4	25	100	100	100	100	100	100	100	3
liver	5.00	2.00	100	100	100	100	6	44	100	100	100	100	100	100	100	10
liver	10.00	1.10	21	26	12	8	5	11	14	50	23	21	21	17	59	
liver	10.00	1.20	89	89	54	23	6	13	66	98	90	85	81	82	99	
liver	10.00	1.50	100	100	100	100	5	37	100	100	100	100	100	100	100	3
liver	10.00	2.00	100	100	100	100	8	67	100	100	100	100	100	100	100	10
liver	25.00	1.10	64	71	36	13	4	12	41	89	67	61	58	53	90	
liver	25.00	1.20	100	100	95	66	5	23	100	100	100	100	100	100	100	
liver	25.00	1.50	100	100	100	100	7	69	100	100	100	100	100	100	100	9
liver	25.00	2.00	100	100	100	100	9	96	100	100	100	100	100	100	100	10
liver	50.00	1.10	89	96	69	28	5	17	82	99	98	94	88	93	100	
liver	50.00	1.20	100	100	100	98	4	38	100	100	100	100	100	100	100	4

County level data																
Site	size	Risk	Besag Newell			normalized	Moran's I		Tango		CEPP			Satscan		UI
			5%	10%	20%		rate	regular	ADJ	PDM	5%	10%	20%	circle	ellipse	
liver	50.00	1.50	100	100	100	100	7	87	100	100	100	100	100	100	100	
liver	50.00	2.00	100	100	100	100	20	100	100	100	100	100	100	100	100	10
lungf	2.50	1.10	8	6	6	5	5	5	5	18	7	7	6	6	28	
lungf	2.50	1.20	19	7	7	6	5	5	8	24	15	12	7	11	35	
lungf	2.50	1.50	91	42	24	21	5	3	37	77	89	61	27	77	93	3
lungf	2.50	2.00	100	99	88	88	5	1	99	100	100	100	88	100	100	9
lungf	5.00	1.10	11	5	6	6	4	4	6	17	11	8	6	8	33	
lungf	5.00	1.20	36	12	9	8	5	4	13	35	37	24	9	23	53	
lungf	5.00	1.50	100	85	51	53	6	2	83	98	100	97	59	99	100	4
lungf	5.00	2.00	100	100	100	100	6	1	100	100	100	100	100	100	100	10
lungf	10.00	1.10	16	7	8	9	6	4	7	24	16	12	8	11	44	
lungf	10.00	1.20	71	19	20	17	6	4	24	58	70	45	18	57	84	
lungf	10.00	1.50	100	98	91	92	5	1	100	100	100	100	92	100	100	3
lungf	10.00	2.00	100	100	100	100	8	1	100	100	100	100	100	100	100	10
lungf	25.00	1.10	31	12	12	9	4	4	16	46	46	27	14	28	64	
lungf	25.00	1.20	99	67	48	40	5	3	68	97	100	94	49	99	100	
lungf	25.00	1.50	100	100	100	100	6	2	100	100	100	100	100	100	100	3
lungf	25.00	2.00	100	100	100	100	12	0	100	100	100	100	100	100	100	10
lungf	50.00	1.10	73	27	23	20	5	3	33	72	87	58	22	70	91	
lungf	50.00	1.20	100	98	87	84	6	2	99	100	100	100	87	100	100	
lungf	50.00	1.50	100	100	100	100	8	1	100	100	100	100	100	100	100	9
lungf	50.00	2.00	100	100	100	100	17	0	100	100	100	100	100	100	100	10
lungm	2.50	1.10	7	7	8	5	5	5	5	19	8	9	11	8	36	
lungm	2.50	1.20	16	23	25	7	5	5	9	35	21	30	35	20	54	
lungm	2.50	1.50	90	93	93	20	5	3	20	95	89	96	98	94	99	3
lungm	2.50	2.00	100	100	100	70	4	2	82	100	100	100	100	100	100	4
lungm	5.00	1.10	12	14	15	6	5	5	5	26	13	21	19	14	46	
lungm	5.00	1.20	36	45	47	7	4	3	9	56	41	56	64	43	75	3
lungm	5.00	1.50	100	100	100	38	6	4	49	100	100	100	100	100	100	10
lungm	5.00	2.00	100	100	100	99	7	3	100	100	100	100	100	100	100	10
lungm	10.00	1.10	22	22	21	7	4	5	6	34	20	30	36	20	57	3
lungm	10.00	1.20	78	80	76	13	6	5	14	84	72	86	93	81	95	3
lungm	10.00	1.50	100	100	100	83	6	4	92	100	100	100	100	100	100	9
lungm	10.00	2.00	100	100	100	100	11	2	100	100	100	100	100	100	100	10
lungm	25.00	1.10	58	57	58	9	6	5	8	68	51	67	74	55	83	3
lungm	25.00	1.20	100	100	100	31	5	5	39	100	100	100	100	100	100	4
lungm	25.00	1.50	100	100	100	100	6	3	100	100	100	100	100	100	100	10
lungm	25.00	2.00	100	100	100	100	15	2	100	100	100	100	100	100	100	10
lungm	50.00	1.10	91	91	88	16	6	5	16	92	84	94	96	90	97	3
lungm	50.00	1.20	100	100	100	70	5	3	82	100	100	100	100	100	100	9

County level data																
Site	size	Risk	Besag Newell				Moran's I			Tango		CEPP			Satscan	
			5%	10%	20%	normalized	rate	regular	ADJ	PDM	5%	10%	20%	circle	ellipse	UI
lungm	50.00	1.50	100	100	100	100	9	3	100	100	100	100	100	100	100	10
lungm	50.00	2.00	100	100	100	100	21	2	100	100	100	100	100	100	100	10
prostate	2.50	1.10	8	7	7	7	5	5	5	18	7	8	7	7	30	10
prostate	2.50	1.20	20	22	17	8	4	4	7	32	19	24	19	15	43	10
prostate	2.50	1.50	90	91	86	26	4	3	29	92	92	94	88	88	98	10
prostate	2.50	2.00	100	100	100	93	6	2	98	100	100	100	100	100	100	9
prostate	5.00	1.10	11	14	10	7	5	4	6	22	12	15	14	9	37	10
prostate	5.00	1.20	38	47	38	10	5	6	11	53	45	49	43	32	68	10
prostate	5.00	1.50	100	100	100	61	5	4	68	100	100	100	99	100	100	10
prostate	5.00	2.00	100	100	100	100	9	2	100	100	100	100	100	100	100	10
prostate	10.00	1.10	21	25	16	8	5	5	7	32	18	23	21	18	51	10
prostate	10.00	1.20	77	80	67	19	5	3	17	80	79	81	72	69	93	10
prostate	10.00	1.50	100	100	100	97	6	3	99	100	100	100	100	100	100	9
prostate	10.00	2.00	100	100	100	100	8	2	100	100	100	100	100	100	100	10
prostate	25.00	1.10	54	59	45	12	5	4	13	62	55	58	49	44	79	10
prostate	25.00	1.20	100	100	100	52	4	5	57	100	100	100	99	100	100	10
prostate	25.00	1.50	100	100	100	100	7	3	100	100	100	100	100	100	100	9
prostate	25.00	2.00	100	100	100	100	16	2	100	100	100	100	100	100	100	10
prostate	50.00	1.10	91	91	80	22	5	4	26	88	90	92	81	84	97	10
prostate	50.00	1.20	100	100	100	93	6	3	97	100	100	100	100	100	100	10
prostate	50.00	1.50	100	100	100	100	9	2	100	100	100	100	100	100	100	10
prostate	50.00	2.00	100	100	100	100	29	1	100	100	100	100	100	100	100	10

Figure 1. Cancer patterns at county level for mortality data from 1998–2002. The values for c is selected to be 0.1, 0.2, 0.5 and 1 in our study. The maximum relative risks in the maps are then 1.1, 1.2, 1.5 and 2.0 with different c 's.

Cervix Cancer Pattern

Colorectal Cancer Pattern

Figure 2. Selected cancer patterns at state level. The relative risk for each state is the weighted average of county level relative risk with county population as weight.

Figure 3.
Powers for selected sites, relative risks (1.1 and 1.2), sample sizes and tests

Figure 4.

Detected clusters for cluster detection methods on Cervix pattern, $RR=1.2$ and sample size 50k. and 300k. For ULS, FSS, LISA, there are 6 levels of colors from light to dark representing clustercount 0, 1–200, 201–400, 401–600, 601–800, 801–1000, which are the numbers of times the cells being counted as in the detected clusters over the 1000 repeats. For SaTScan methods, only one random data set from the 1000 repeats is evaluated, so there are two colors (light for 0 and dark for 1) in their maps, 0 implies outside the detected clusters and 1 otherwise.

Figure 5.

Detected clusters for cluster detection methods on Colorectal pattern, $RR=1.2$ and sample size 50k and 300k. For ULS, FSS, LISA, there are 6 levels of colors from light to dark representing clustercount 0, 1–200, 201–400, 401–600, 601–800, 801–1000, which are the numbers of times the cells being counted as in the detected clusters over the 1000 repeats. For SaTScan methods, only one random data set from the 1000 repeats is evaluated, so there are two colors (light for 0 and dark for 1) in their maps, 0 implies outside the detected clusters and 1 otherwise.

Table 1

Characteristics of the eight selected cancer patterns (sites). *c* is selected to be 0.1, 0.2, 0.5 and 1.0 in our study. We call the areas with higher relative risks (relative risk > 1) cluster areas. “stdpop” is the standard deviation of the county population in the cluster areas. The population is from 2000 census county level population. “lungm” is lung male and “lungf” is lung female. Average Beale is the average of Beale codes at county level from census. The values of the Beale code changes from 1 to 9, with small values representing for counties in metro area with varying high population, moderate values for non-metro counties with varying population and adjacent to a metro area, large values for non-metro counties completely rural or with low populations, not adjacent to a metro area. Percent population is the percent of population in the cluster areas over the total population. Percent expected cases is the percent of expected cases in cluster the areas over the total expected cases.

Site	relative risk in cluster regions	# county	average pop (k)	stdpop (k)	average Beale codes	percent pop (%)	percent expected cases (%)
bladder	1+c	356	127.38	224.68	4.13	16.22	18.85
cervix	1+c	650	58.60	174.99	4.91	13.62	15.91
colorectal	1+c	378	73.20	133.91	4.70	9.90	11.65
kidney	1+0.5c	167	91.97	337.00	5.03	5.49	5.85
	1+c	531	70.46	268.97	5.25	13.38	15.56
	1+0.5c or 1+c	698	75.60	286.60	5.20	18.87	21.41
liver	1+c	495	90.53	257.59	5.11	16.03	18.64
lungf	1+c	324	76.53	170.82	4.80	8.87	10.46
lungm	1+0.5c	714	61.55	114.87	4.86	15.72	16.92
	1+c	212	38.72	61.04	5.31	2.94	3.45
	1+0.5c or 1+c	926	56.32	105.42	4.96	18.66	20.37
prostate	1+c	529	50.81	94.70	5.09	9.61	11.32

Table 2

The powers in this table is the average power over the powers by site, and by size and relative risk over the 160 combinations. The bold numbers are the places that the particular method begins to have average power bigger than 90%.

Site	County level data										State level data			
	Besag Newell					Moran's I		Tango		CEPP		Satscan		FSS
	5%	10%	20%	normalized	rate	regular	ADJ	PDM	5%	10%	20%	circle	ellipse	
bladder	84.1	81.4	78.3	63.2	6.5	32.7	74.0	88.5	82.3	83.1	79.3	80.7	89.4	80.1
cervix	78.0	80.9	76.2	58.3	8.7	0.9	66.4	83.7	77.6	78.3	80.7	76.7	87.9	76.3
colorectal	65.9	61.3	57.9	54.4	6.4	9.9	58.3	70.4	66.7	67.3	65.4	69.5	82.9	67.3
kidney	80.0	83.3	80.8	58.3	8.1	0.6	72.9	87.0	77.8	80.9	82.4	75.0	86.3	83.9
liver	77.3	79.3	70.0	58.6	6.2	36.0	72.5	85.9	78.7	77.5	76.3	75.4	87.2	75.7
lungf	72.8	59.1	53.9	52.8	6.4	2.5	59.9	73.4	73.8	67.2	54.4	69.4	81.2	65.4
lungm	75.5	76.7	76.5	49.0	6.9	3.7	51.8	80.4	74.9	79.5	81.3	76.3	87.1	76.6
prostate	75.5	76.8	73.2	55.7	7.3	3.4	56.9	78.8	75.9	77.2	74.6	73.2	84.8	74.5
Risk size (k)														
1.10	2.50	8.9	8.9	7.5	5.7	4.9	5.0	5.8	20.7	7.8	8.8	8.5	7.1	8.4
1.10	5.00	12.2	14.1	11.9	6.9	4.9	4.7	7.7	27.1	13.8	15.5	14.9	10.7	13.9
1.10	10.00	22.6	24.0	17.4	7.6	5.0	5.3	10.2	40.6	20.8	24.6	24.9	18.5	20.1
1.10	25.00	56.8	56.1	45.4	12.0	4.6	5.3	25.9	71.2	56.8	58.1	55.6	49.0	54.7
1.10	50.00	85.9	79.6	71.3	25.4	5.7	5.9	51.5	88.7	88.2	85.9	78.6	85.1	82.3
1.20	2.50	23.5	23.9	18.8	7.4	4.5	4.7	10.8	40.2	20.8	23.8	22.9	17.3	20.9
1.20	5.00	44.6	47.1	37.2	10.3	4.4	5.1	20.7	61.1	45.4	49.0	46.2	37.2	44.0
1.20	10.00	79.3	72.2	61.5	21.0	5.6	5.5	40.1	82.7	77.4	76.9	72.2	74.7	72.2
1.20	25.00	98.6	92.0	87.6	56.2	5.1	7.4	77.7	98.0	98.6	98.1	91.4	99.3	97.8
1.20	50.00	100	99.6	97.3	91.5	5.3	9.5	96.9	100	100	100	98.3	100	100
1.50	2.50	91.0	83.5	75.5	29.6	4.9	5.6	55.0	91.0	90.1	87.7	82.1	92.0	86.1
1.50	5.00	99.0	96.5	89.6	64.9	5.5	7.3	83.7	99.0	99.4	99.1	93.4	99.8	98.9
1.50	10.00	100	99.6	98.3	95.6	6.0	10.2	98.8	100	100	100	98.9	100	100
1.50	25.00	100	100	100	100	6.5	17.6	100	100	100	100	100	100	100
1.50	50.00	100	100	100	100	8.6	23.7	100	100	100	100	100	100	100

Site	County level data										State level data					
	Besag Newell			Moran's I		Tango			CEPP		Satscan		ULS	ULS	FSS	
	5%	10%	20%	normalized	rate	regular	ADJ	PDM	5%	10%	20%	circle	ellipse			
2.00	2.50	100	99.8	97.7	91.8	6.0	9.0	97.3	100	100	98.4	100	100	91.6	99.5	100
2.00	5.00	100	100	99.9	99.9	7.3	14.2	100	100	100	100	100	100	99.9	100	100
2.00	10.00	100	100	100	100	9.3	20.5	100	100	100	100	100	100	100	100	100
2.00	25.00	100	100	100	100	13.5	27.8	100	100	100	100	100	100	100	100	100
2.00	50.00	100	100	100	100	23.2	29.8	100	100	100	100	100	100	100	100	100

Table 3

Precision ($\bar{r}_T$, $\bar{r}_D$ and related 95% confidence interval) of cluster detection methods for data with $c=0.2$ (relative risk=1.2), and sample size 5k and 50k. SaTScanO and SaTScanE only have r_T and r_D for a data set randomly selected out of the 1000 replicates under each H_a .

size	site	CEPP			Satscan		LISA		ULS		FSSstate
		5%	10%	20%	O	E	nor	reg	county	state	
r_T											
5	bladder	0.20 (0.00, 0.49)	0.37 (0.00, 0.83)	0.38 (0.00, 1.00)	0.00	0.00	0.10 (0.05, 0.14)	0.09 (0.04, 0.14)	0.15 (0.00, 0.49)	0.25 (0.00, 0.88)	0.21 (0.00, 0.63)
5	cervix	0.12 (0.00, 0.39)	0.22 (0.00, 0.64)	0.46 (0.00, 1.00)	0.00	0.00	0.09 (0.06, 0.13)	0.05 (0.00, 0.11)	0.01 (0.00, 0.08)	0.24 (0.00, 0.86)	0.20 (0.00, 0.68)
5	colorectal	0.04 (0.00, 0.19)	0.07 (0.00, 0.33)	0.09 (0.00, 0.43)	0.00	0.00	0.09 (0.05, 0.13)	0.06 (0.01, 0.11)	0.04 (0.00, 0.20)	0.13 (0.00, 0.62)	0.08 (0.00, 0.38)
5	kidney	0.09 (0.00, 0.30)	0.20 (0.00, 0.55)	0.40 (0.00, 1.00)	0.00	0.00	0.13 (0.09, 0.16)	0.05 (0.00, 0.11)	0.01 (0.00, 0.06)	0.30 (0.00, 0.91)	0.35 (0.00, 0.80)
5	liver	0.14 (0.00, 0.44)	0.19 (0.00, 0.60)	0.25 (0.00, 0.84)	0.00	0.00	0.11 (0.07, 0.15)	0.06 (0.00, 0.12)	0.00 (0.00, 0.03)	0.20 (0.00, 0.80)	0.16 (0.00, 0.56)
5	lungf	0.12 (0.00, 0.47)	0.09 (0.00, 0.41)	0.04 (0.00, 0.26)	0.00	0.00	0.10 (0.06, 0.15)	0.06 (0.01, 0.12)	0.01 (0.00, 0.09)	0.08 (0.00, 0.46)	0.05 (0.00, 0.27)
5	lungm	0.11 (0.00, 0.39)	0.27 (0.00, 0.77)	0.52 (0.00, 1.00)	0.00	0.00	0.08 (0.05, 0.10)	0.06 (0.00, 0.11)	0.03 (0.00, 0.14)	0.23 (0.00, 0.82)	0.20 (0.00, 0.67)
5	prostate	0.15 (0.00, 0.50)	0.23 (0.00, 0.71)	0.27 (0.00, 0.90)	0.02	0.08	0.10 (0.06, 0.14)	0.06 (0.00, 0.13)	0.02 (0.00, 0.11)	0.15 (0.00, 0.65)	0.11 (0.00, 0.39)
	average	0.12 (0.00, 0.40)	0.20 (0.00, 0.61)	0.30 (0.00, 0.80)	0.00	0.01	0.10 (0.06, 0.14)	0.06 (0.01, 0.12)	0.03 (0.00, 0.15)	0.20 (0.00, 0.75)	0.17 (0.00, 0.55)
50	bladder	0.41 (0.25, 0.57)	0.56 (0.45, 0.67)	0.76 (0.71, 0.82)	0.69	0.86	0.28 (0.22, 0.35)	0.19 (0.15, 0.23)	0.57 (0.13, 1.00)	0.75 (0.52, 0.99)	0.34 (0.00, 0.73)
50	cervix	0.55 (0.25, 0.85)	0.48 (0.33, 0.62)	0.81 (0.74, 0.88)	0.86	0.82	0.17 (0.13, 0.22)	0.06 (0.04, 0.08)	0.49 (0.10, 0.87)	0.64 (0.44, 0.84)	0.52 (0.41, 0.63)
50	colorectal	0.39 (0.09, 0.68)	0.40 (0.11, 0.70)	0.47 (0.32, 0.61)	0.65	0.77	0.20 (0.15, 0.26)	0.09 (0.06, 0.12)	0.28 (0.00, 0.84)	0.56 (0.22, 0.90)	0.40 (0.24, 0.55)
50	kidney	0.25 (0.03, 0.47)	0.37 (0.08, 0.66)	0.66 (0.47, 0.85)	0.89	0.42	0.15 (0.11, 0.19)	0.06 (0.04, 0.08)	0.33 (0.00, 0.77)	0.66 (0.45, 0.87)	0.55 (0.38, 0.71)
50	liver	0.51 (0.33, 0.69)	0.49 (0.24, 0.75)	0.63 (0.47, 0.78)	0.77	0.94	0.19 (0.14, 0.23)	0.09 (0.06, 0.12)	0.25 (0.00, 0.78)	0.67 (0.48, 0.86)	0.42 (0.25, 0.59)
50	lungf	0.70 (0.41, 0.99)	0.59 (0.20, 0.98)	0.46 (0.00, 0.93)	0.69	0.69	0.21 (0.15, 0.28)	0.10 (0.07, 0.13)	0.05 (0.00, 0.32)	0.41 (0.02, 0.79)	0.30 (0.10, 0.51)
50	lungm	0.32 (0.16, 0.47)	0.55 (0.47, 0.64)	0.95 (0.90, 1.00)	0.97	0.96	0.11 (0.07, 0.14)	0.06 (0.03, 0.08)	0.50 (0.30, 0.71)	0.60 (0.34, 0.86)	0.56 (0.33, 0.79)
50	prostate	0.43 (0.33, 0.52)	0.52 (0.45, 0.60)	0.70 (0.59, 0.82)	0.74	0.59	0.18 (0.13, 0.23)	0.07 (0.04, 0.09)	0.34 (0.00, 0.81)	0.34 (0.04, 0.63)	0.28 (0.16, 0.40)
	average	0.44 (0.23, 0.66)	0.50 (0.29, 0.70)	0.68 (0.52, 0.84)	0.78	0.76	0.19 (0.14, 0.24)	0.09 (0.06, 0.12)	0.35 (0.07, 0.76)	0.58 (0.31, 0.84)	0.42 (0.24, 0.61)
r_D											
5	bladder	0.59 (0.00, 1.00)	0.51 (0.00, 1.00)	0.24 (0.00, 0.67)	0.00	0.00	0.11 (0.06, 0.15)	0.16 (0.03, 0.29)	0.31 (0.00, 1.00)	0.30 (0.00, 1.00)	0.51 (0.00, 1.00)
5	cervix	0.33 (0.00, 1.00)	0.29 (0.00, 0.86)	0.32 (0.00, 0.82)	0.00	0.00	0.18 (0.13, 0.24)	0.19 (0.09, 0.30)	0.04 (0.00, 0.32)	0.22 (0.00, 0.80)	0.37 (0.00, 1.00)
5	colorectal	0.10 (0.00, 0.51)	0.10 (0.00, 0.44)	0.06 (0.00, 0.28)	0.00	0.00	0.11 (0.06, 0.15)	0.12 (0.03, 0.22)	0.10 (0.00, 0.45)	0.12 (0.00, 0.60)	0.18 (0.00, 0.86)
5	kidney	0.30 (0.00, 0.98)	0.31 (0.00, 0.80)	0.27 (0.00, 0.65)	0.00	0.00	0.28 (0.22, 0.34)	0.22 (0.11, 0.33)	0.04 (0.00, 0.28)	0.37 (0.00, 1.00)	0.65 (0.00, 1.00)
5	liver	0.32 (0.00, 0.96)	0.23 (0.00, 0.72)	0.14 (0.00, 0.48)	0.00	0.00	0.17 (0.11, 0.22)	0.17 (0.08, 0.27)	0.01 (0.00, 0.11)	0.19 (0.00, 0.78)	0.28 (0.00, 1.00)

size	site	CEPP			Satscan		LISA		ULS		FSSstate
		5%	10%	20%	O	E	nor	reg	county	state	
5	lungf	0.21 (0.00, 0.82)	0.09 (0.00, 0.44)	0.02 (0.00, 0.13)	0.00	0.00	0.10 (0.06, 0.15)	0.12 (0.03, 0.20)	0.03 (0.00, 0.19)	0.07 (0.00, 0.44)	0.12 (0.00, 0.62)
5	lungm	0.37 (0.00, 1.00)	0.50 (0.00, 1.00)	0.52 (0.00, 1.00)	0.00	0.00	0.22 (0.16, 0.28)	0.27 (0.14, 0.39)	0.16 (0.00, 0.78)	0.23 (0.00, 0.84)	0.38 (0.00, 1.00)
5	prostate	0.34 (0.00, 1.00)	0.27 (0.00, 0.86)	0.17 (0.00, 0.57)	0.92	0.98	0.16 (0.11, 0.22)	0.15 (0.07, 0.23)	0.06 (0.00, 0.42)	0.18 (0.00, 0.75)	0.33 (0.00, 1.00)
	average	0.32 (0.00, 0.91)	0.29 (0.00, 0.76)	0.22 (0.00, 0.58)	0.12	0.12	0.17 (0.11, 0.22)	0.17 (0.07, 0.28)	0.09 (0.00, 0.44)	0.21 (0.00, 0.78)	0.35 (0.00, 0.94)
50	bladder	0.89 (0.79, 0.99)	0.78 (0.70, 0.86)	0.48 (0.41, 0.55)	0.77	0.52	0.39 (0.33, 0.46)	0.19 (0.12, 0.27)	0.46 (0.01, 0.91)	0.93 (0.74, 1.00)	0.99 (0.94, 1.00)
50	cervix	0.76 (0.64, 0.89)	0.64 (0.53, 0.75)	0.55 (0.51, 0.59)	0.50	0.69	0.45 (0.36, 0.53)	0.15 (0.09, 0.22)	0.43 (0.07, 0.79)	0.89 (0.63, 1.00)	0.99 (0.91, 1.00)
50	colorectal	0.50 (0.31, 0.69)	0.47 (0.26, 0.67)	0.31 (0.21, 0.42)	0.45	0.47	0.34 (0.26, 0.43)	0.13 (0.07, 0.19)	0.15 (0.00, 0.45)	0.79 (0.44, 1.00)	0.92 (0.74, 1.00)
50	kidney	0.88 (0.64, 1.00)	0.56 (0.45, 0.68)	0.44 (0.39, 0.48)	0.41	0.60	0.43 (0.35, 0.51)	0.25 (0.18, 0.32)	0.34 (0.00, 0.77)	0.93 (0.76, 1.00)	0.97 (0.88, 1.00)
50	liver	0.67 (0.57, 0.78)	0.49 (0.36, 0.62)	0.36 (0.28, 0.45)	0.73	0.63	0.38 (0.31, 0.46)	0.20 (0.13, 0.27)	0.20 (0.00, 0.65)	0.89 (0.62, 1.00)	0.95 (0.74, 1.00)
50	lungf	0.67 (0.54, 0.80)	0.42 (0.22, 0.62)	0.18 (0.00, 0.40)	0.87	0.87	0.32 (0.23, 0.40)	0.13 (0.07, 0.19)	0.02 (0.00, 0.14)	0.80 (0.39, 1.00)	0.88 (0.62, 1.00)
50	lungm	0.99 (0.92, 1.00)	0.98 (0.93, 1.00)	0.91 (0.82, 1.00)	0.87	0.89	0.45 (0.36, 0.55)	0.19 (0.12, 0.26)	0.61 (0.32, 0.90)	0.89 (0.62, 1.00)	1.00 (0.93, 1.00)
50	prostate	0.91 (0.79, 1.00)	0.61 (0.53, 0.69)	0.45 (0.33, 0.57)	0.70	0.74	0.40 (0.32, 0.49)	0.13 (0.08, 0.18)	0.28 (0.00, 0.70)	0.92 (0.68, 1.00)	0.98 (0.86, 1.00)
	average	0.79 (0.65, 0.90)	0.62 (0.50, 0.74)	0.46 (0.37, 0.56)	0.66	0.68	0.40 (0.31, 0.48)	0.17 (0.11, 0.24)	0.31 (0.05, 0.66)	0.88 (0.61, 1.00)	0.96 (0.83, 1.00)