STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:853–865
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2418

The geography of power: Statistical performance of tests of clusters and clustering in heterogeneous populations

Lance A. Waller[1,*,†], Elizabeth G. Hill[2] and Rose Anne Rudd[1]

[1]*Department of Biostatistics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, U.S.A.*
[2]*Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC, U.S.A.*

SUMMARY

Heterogeneous population densities complicate comparisons of statistical power between hypothesis tests evaluating spatial clusters or clustering of disease. Specifically, the *location* of a cluster within a heterogeneously distributed population at risk impacts power properties, complicating comparisons of tests, and allowing one to map spatial variations in statistical power for different tests. Such maps provide insight into the overall power of a particular test, and also indicate areas within the study area where tests are more or less likely to detect the same local increase in relative risk. While such maps are largely driven by local sample size, we also find differences due to features of the statistics themselves. We illustrate these concepts using two tests: Tango's index of clustering and the spatial scan statistic. Furthermore, assessments of the accuracy of the 'most likely cluster' involve not only statistical power, but also spatial accuracy in identifying the location of a true underlying cluster. We illustrate these concepts via induction of artificial clusters within the observed incidence of severe cardiac birth defects in Santa Clara County, CA in 1981. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: disease clusters; disease clustering; power; spatial scan statistic; spatial patterns of disease; medical geography

1. INTRODUCTION

The geographic distribution of disease is often of interest to epidemiologists and other medical researchers because the spatial distribution may identify areas of raised incidence requiring targeted interventions or even suggest causal determinants of disease. In addition, spatial

*Correspondence to: Lance A. Waller, Department of Biostatistics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, U.S.A.
†E-mail: lwaller@sph.emory.edu

Contract/grant sponsor: Associations of Schools of Public Health/Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry Cooperative Agreement Project; contract/grant number: S779-18/18
Contract/grant sponsor: National Institute of Environmental Health Sciences; contract/grant number: R01-ES007750

Received 23 August 2005
Accepted 1 September 2005

patterns in incidence are of interest to the general public, particularly regarding possible relationships between local disease incidence and local environmental stressors (e.g. hazardous waste sites, cellular telephone towers, high-voltage electrical lines, or other perceived sources of hazard).

Assessments of geographical patterns of disease often involve statistical tests to detect unusual spatial (or spatio-temporal) aggregations of incident cases of disease, or disease 'clusters', and the statistical, epidemiological, and geographic literature contain many proposed methods for identifying clusters and drawing inference on them. Elliott *et al*. [1], Tango [2], Wakefield *et al*. [3], Kulldorff [4] and Waller and Gotway [5] all provide recent reviews of many such methods.

Before considering particular tests, we first note that not all tests seek to detect the same sort of deviations from a conceptual null hypothesis of 'no clustering'. That is, not all tests consider the same sorts of alternative hypotheses. Besag and Newell [6] and Kulldorff and Nagarwalla [7] summarize helpful terminology categorizing many of the hypothesis tests proposed in the literature by their associated conceptual alternative hypotheses. To borrow the terminology of Besag and Newell [6], 'tests of disease *clustering*' examine the overall tendency of disease cases to cluster together across a region, while 'tests to detect *clusters*' seek to define areas most likely experiencing increased incidence rates. In practice, the distinction often presents via a single significance value summarizing patterns across the entire data set for statistics assessing (global) clustering, and cluster-specific significance values associated with particular clusters (e.g. the most suspicious collection of cases) for tests to detect disease clusters.

The number of tests of clustering and to detect clusters proposed raises the issue of how to compare performance across methods. In general, one typically compares competing hypothesis tests via power (the probability of rejecting the null hypothesis when an alternative is true) or similar measures. The literature contains several power comparisons of proposed tests of clustering or tests to detect clusters, typically compared as functions of increasing relative risk within the cluster. Relevant to the discussion below, some comparisons report different power results for clusters located in different parts of the study area, in particular, both Hill *et al*. [8] and Gangnon and Clayton [9] note that the power of the tests considered in their papers varies across a limited number of locations for a disease cluster within the geographic region under study.

Based on such previous considerations, we expand the discussion regarding statistical power and performance to an explicitly spatial setting, mapping power and related quantities based on the location of a single cluster within an example data set (the Santa Clara county severe cardiac birth defect considered by Shaw *et al.* [10] and Hill *et al.* [8]). The example reveals the impact of local geography (in particular, population density) on power comparisons between statistical tests of spatial pattern. While based on a particular data example, the implications of these illustrations extend beyond the particular data set and tests considered here to any sort of power comparison of tests to detect clustering or clusters when applied within a heterogeneously distributed population at risk, and raise important issues regarding assessments of statistical performance for both tests of clustering and tests to detect clusters.

We consider alternatives defined by a single disease cluster associated with a fixed increase in relative risk centred at each of 259 locations, and assess both the power to identify the cluster and the power to detect clustering generated by this single cluster, as a function of the location of the cluster. To illustrate the concepts outlined above, we present maps depicting

the spatial variation in the statistical power based on the Santa Clara County cardiac defect data for two specific hypothesis tests, namely Kulldorff's [11] spatial scan statistic (a popular test to detect clusters), and Tango's [12] statistic (a test of clustering) based on an application to the Santa Clara County cardiac defect data.

In the following sections we briefly describe the example data set, define the tests under consideration, and outline our approach. We conclude with results and discussion of the spatially varying performance of such tests in both particular and general terms.

2. METHODS

2.1. Data description

The data involve major cardiac anomalies in live births in 1981 for Santa Clara County, CA and provide an illustrative example of many analytic assessments of disease clusters and/or clustering. Retrospective data collection followed high rates of adverse pregnancy outcomes in 1982 reported by citizens in a collection of seven contiguous census tracts serviced by two drinking water wells contaminated with organic solvents. The data initially included 20 886 live births geocoded to 294 census tracts of residency within the county. We follow Hill *et al.* [8] and exclude thirty-five of the 294 census tracts from analysis due to mismatched identification numbers between the data set and the 1980 United States census tracts. The thirty-five excluded census tracts contain only 87 live births (0.42 per cent of the total) and no cases of major cardiac anomalies in live births. The 259 census tracts remaining for analysis contain 20 799 live births, with 71 experiencing severe cardiac defects. For the purposes of our analysis (and similar to many such studies in the literature) we locate all births and cases at the centroid of the census tract of residence, due to confidentiality restrictions.

2.2. Hypothesis formulation

We consider a null hypothesis where each birth is equally likely to experience the adverse outcome, so the expected incidence count of each area varies only due to heterogeneous population sizes (here, number of births).

To formalize, suppose we have a geographic study region partitioned into I tracts, where n_i denotes the number of persons at risk (live births) in tract i, $i=1,\ldots,I$, and $n_+=\sum_{i=1}^{I} n_i$ is the total number of persons at risk in the study area. Consider the number of incident cases of disease in tract i as a random variable C_i with observed value c_i, and denote the total number of observed cases by $c_+=\sum_{i=1}^{I} c_i$. We define a null hypothesis of no clustering by

$$H_0:\ \{C_i\} \text{ are independent Poisson random variables with } E(C_i)=\lambda n_i,\quad i=1,\ldots,I \tag{1}$$

where λ is the baseline incidence rate. That is, under H_0, the expected number of cases in each tract is the baseline rate of disease multiplied by the number at risk in tract i. When the baseline incidence rate is unknown, one obtains the null distribution of the $\{C_i\}$ by conditioning on the sufficient statistic c_+, defining a conditional null hypothesis as

$$H_0:\ C_1,\ldots,C_I \mid c_+ \sim \text{multinomial}(c_+, n_1/n_+,\ldots,n_I/n_+) \tag{2}$$

The alternative hypothesis depends on the model of clustering the investigator wishes to detect (Waller and Jacquez [13]). For this study we consider an alternative hypothesis defined by

$$H_1: E(C_i) = \lambda n_i(1 + \delta_i \varepsilon) \tag{3}$$

where

$$\delta_i = \begin{cases} 1 & \text{if tract } i \text{ is in the cluster} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$\varepsilon = \mathrm{RR} - 1$ for $i = 1, \ldots, I$, and RR is the relative risk of disease within the cluster [6]. For unknown λ we consider the conditional alternative hypothesis

$$H_1: C_1, \ldots, C_I \mid c_+ \sim \text{multinomial}(c_+, \pi_1, \ldots, \pi_I) \tag{5}$$

where

$$\pi_i = \frac{n_i(1 + \delta_i \varepsilon)}{\sum_{i=1}^{I} n_i(1 + \delta_i \varepsilon)}, \quad i = 1, \ldots, I \tag{6}$$

2.3. Kulldorff's spatial scan statistic

Kulldorff [11] defines a spatial scan statistic to detect the most likely cluster based on locally observed likelihood ratio test statistics. The scan statistic is constructed based on circles of increasing radii centred at each tract centroid until a maximum radius, selected *a priori*, is attained. Kulldorff and Nagarwalla [7] suggest a largest circle containing 50 per cent of the total at-risk population, and we use this definition here. Kulldorff [11] refers to the collection of tracts having their centroid contained within each circle as a zone, z. The scan statistic considers each zone as a potential cluster, and identifies the zones least consistent with the null hypothesis. We note that for data consisting of regional counts, while the centroids are contained within a circle, zones actually reflect an irregular area defined by the union of polygonal regions associated with the tracts. Let n_z and c_z be the population size and case count, respectively, in zone z. Define p_z and $p_{\bar{z}}$ as the probability of being a case inside and outside zone z, respectively. Based on the null hypothesis $H_0: p_z = p_{\bar{z}}$ *versus* the alternative $H_A: p_z > p_{\bar{z}}$, Kulldorff [11] defines a likelihood ratio statistic proportional to

$$L_z = \left(\frac{c_z}{\hat{\lambda} n_z}\right)^{c_z} \left(\frac{c_+ - c_z}{c_+ - \hat{\lambda} n_z}\right)^{c_+ - c_z} 1[c_z > \hat{\lambda} n_z]$$

where $\hat{\lambda} = c_+/n_+$ is the estimated baseline incidence rate, and $1[c_z > \hat{\lambda} n_z]$ is an indicator function equal to 1 when the number of observed cases in zone z exceeds that expected under H_0, and is equal to 0 otherwise. The most likely cluster is defined by the zone, $\tilde{z}$, maximizing L_z over all possible zones considered.

The statistical significance of $L_{\max} = L_{\tilde{z}}$ is obtained via Monte Carlo simulation. Specifically, the c_+ cases are distributed uniformly among the n_+ individuals according to (2), and the maximum value of L_z is calculated for each simulated data set. The p-value associated with the most likely cluster is the proportion of observed and simulated statistics greater than or

equal to the value of $L_{\max}$ observed in the data. Note that the Monte Carlo inference ranks the observed maximum likelihood ratio statistic $L_{\max}$ from the data among a set comprised of the maximum likelihood ratio statistic from each simulated data set, and not among the statistics observed at the same zone as the maximum in the data set. As a result, inference is not based on the distribution of a likelihood ratio for a particular zone, but rather on the distribution of the maximized likelihood ratio under the null hypothesis, regardless of which zone contains the maximum.

2.4. Tango's general test

Tango [12] considered assessments of spatial clustering via the statistic

$$C_G = (\mathbf{r} - \mathbf{p})^{\mathrm{t}} \mathbf{A} (\mathbf{r} - \mathbf{p}) \tag{7}$$

where $\mathbf{r}^{\mathrm{T}} = (c_1, \ldots, c_I)/c_+$ is an I-dimensional vector of case proportions and $\mathbf{p} = E(\mathbf{r} \mid H_0)$ is an I-dimensional vector of population proportions $p_i = n_i/n_+$, expected under the null hypothesis. The matrix $\mathbf{A} = (a_{ij})$ is an $I \times I$ matrix defining a measure of closeness between any two tracts. We follow Tango [12] and set $a_{ij} = \exp(-d_{ij}/\tau)$, where d_{ij} is the Euclidean distance between the centroids of tracts i and j, and τ a scale parameter. (We note that in our implementation, we use great-circle rather than planar distances for comparability with Kulldorff's [11] spatial scan statistic.) Although large τ will give a test sensitive to large clusters and small τ to small clusters, and Tango [12] claims the choice of τ is unlikely to drastically change the test results, something we investigate with respect to power in the results below. The statistic C_G shows a substantial amount of positive skewness under the null hypothesis, so a normal approximation based on the standardized statistic

$$T_G = \frac{C_G - E(C_G)}{\sqrt{\mathrm{var}(C_G)}} \tag{8}$$

is often inappropriate [12]. Instead, Tango [12] considers the adjusted statistic

$$C_G^* = v + T_G \sqrt{2v} \overset{a}{\sim} \chi_v^2 \tag{9}$$

where

$$v = \frac{8}{\sqrt{\beta_1(C_G)}^2} \tag{10}$$

and $\sqrt{\beta_1(C_G)}$ denotes the skewness of C_G.

To implement the test of Tango [12], one must first define the scale parameter τ. For illustration, we choose τ assigning non-negligible weight to centroids in the seven nearest neighbours. The maximum (great circle) distance between centroids is 7.577km and we choose a value of $\tau = 2.529$ to give a weight of 0.05 to this most distant nearest neighbour. To explore the sensitivity of power to the choice of τ we also consider values $\tau = 1.25, 5.0$, and 7.5.

2.5. Simulation of cases and power of the tests

Through simulation, we show the power of disease clustering tests is not only a function of the strength of clustering (i.e. the increase in relative risk within the cluster) but also the

location of the cluster within the study area. Paralleling the original data setting of a seven-tract area reflecting the putative disease cluster [14], we consider 259 clusters, one centred at each census tract including that tract and its six nearest neighbouring tracts within Santa Clara County. For each of these 259 clusters, we repeatedly assign the 71 observed cases among the 259 tracts under the alternative hypothesis following the conditional multinomial distribution of equations (5) and (6), where in equation (4), $\delta_i = 1$ for each of the seven census tracts defining the cluster. We conduct simulations based on a risk three times higher for individuals inside the cluster compared to those outside the cluster, i.e. we use equation (6) with $\varepsilon = 2$. For simplicity, we make no adjustment of the data set for possible confounding factors.

To explore the power of the tests, we simulate 1000 data sets for each of the 259 cluster locations. For each simulation, we record Kulldorff's [11] $L_{\max}$ and Tango's [12] C_G^*.

The power of Tango's [12] general statistic is simply the proportion of test statistics C_G^* for the clustered data exceeding the 95th percentile of the asymptotic chi-square distribution in (9). We compared results to a Monte Carlo critical value (95th percentile of 1000 data sets simulated under the null hypothesis) with little appreciable difference.

Assessing the power of the spatial scan statistic requires some additional thought. Not only do we wish to determine the proportion of times the null hypothesis is rejected, we also wish to assess the accuracy of the location of the most likely cluster. To begin, we first use the freely available software SaTScan [15] to generate the critical value of $L_{\max}$ for tests conducted at level 0.05. Next, we define power as the proportion of $L_{\max}$ values based on simulations under the alternative that exceed the critical value for which the cluster centre (the tract whose nearest neighbours define the cluster) is identified among any of the tracts within the most likely cluster. We note that other interesting measures of the test's performance include, for example, the proportion of rejections for which the cluster centre is statistically significant and correctly identified, or the proportion of simulations for which the cluster centre is identified anywhere within the most likely cluster regardless of whether or not the null hypothesis was rejected. We explore each of these in Section 3.

3. RESULTS

Figure 1 depicts the number of live births in 1980 for the 259 census tracts of Santa Clara County included in the analysis. The City of San Jose covers much of the northwestern quarter of the county, while the coastal mountain range of California extends in a northwest to southeast direction in a midline through the county. The population of the county is sparsely distributed east of the mountain range and, as a result, the large eastern tract is among those omitted from the analysis.

Figure 2 illustrates the observed geographic distribution of cases (top), and an example of the types of clusters simulated. The next map illustrates one of the 259 clusters considered. The cluster consists of seven tracts. The bottom row of Figure 2 illustrates three simulated data sets based on redistributing the 71 cases among the tracts with increased risk in the cluster tracts. Note that even with the risk approximately tripled in the cluster, the outcome is rare enough that the resulting increase in incidence is subtle and not detectable through simple visual assessment. In other words, while the cluster contains an appreciable increase in multiplicative risk, the outcome is so rare that the additional number of observed cases within the cluster is quite small.

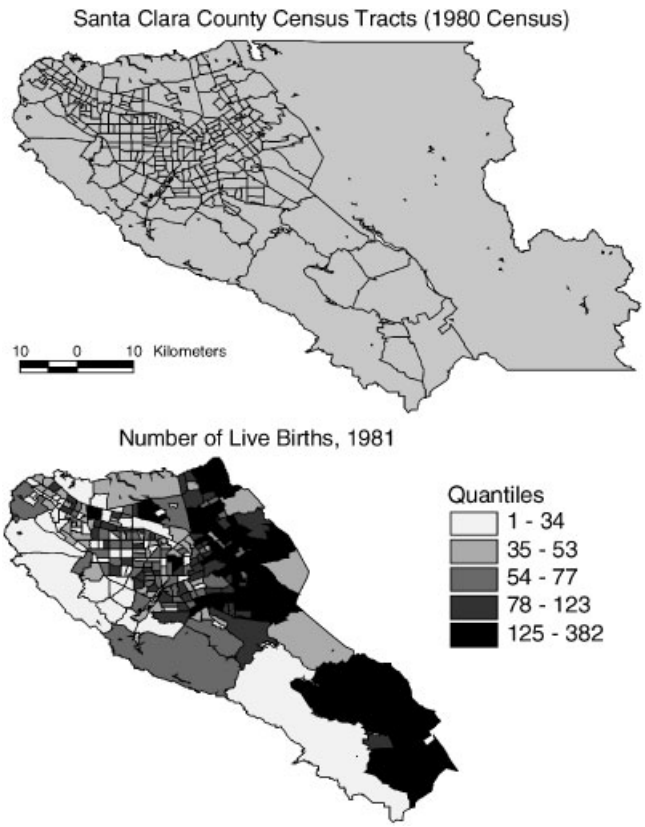

Figure 1. Map of census tracts (1980 U.S. Census) for Santa Clara County (top) and choropleth map of the number of live births per census tract for Santa Clara County, CA in 1981 (bottom).

Since the definition of power is streamlined in the case of a test of clustering, we begin with the results from Tango's [12] test. Figure 3 depicts power maps of Tango's [12] test for $\tau = 1.25, 2.529, 5.0$, and 7.5. Each tract is shaded according to the power of the test to detect general clustering based on a cluster centred at that particular tract and including its six nearest neighbours. Several patterns emerge. First, we note a general lack of power to detect a single cluster of tripled relative risk as evidence of overall clustering, partially due to the rarity of the outcome as noted above. Second, we note considerable spatial variation in power and different geographic variations for different levels of the scale parameter τ. Figure 3 also includes two concentric circles for each map. The inner circle represents the radius within which the measure of closeness is a_{ij} greater than 0.5, and the outer circle represents the radius within which a_{ij} greater than 0.05. The circles illustrate the range of highest values for the proximity weights defining Tango's statistic and we see an impact on patterns of power. Comparing to Figure 1, we see some impact of the geographic distribution of the population at risk (live births) but this influence is tempered by the choice of the scale parameter. For instance, setting $\tau = 1.25$ provides sufficient spatial weighting to capture clusters defined on small tracts (the central portion of the county) with power comparable to the local sample

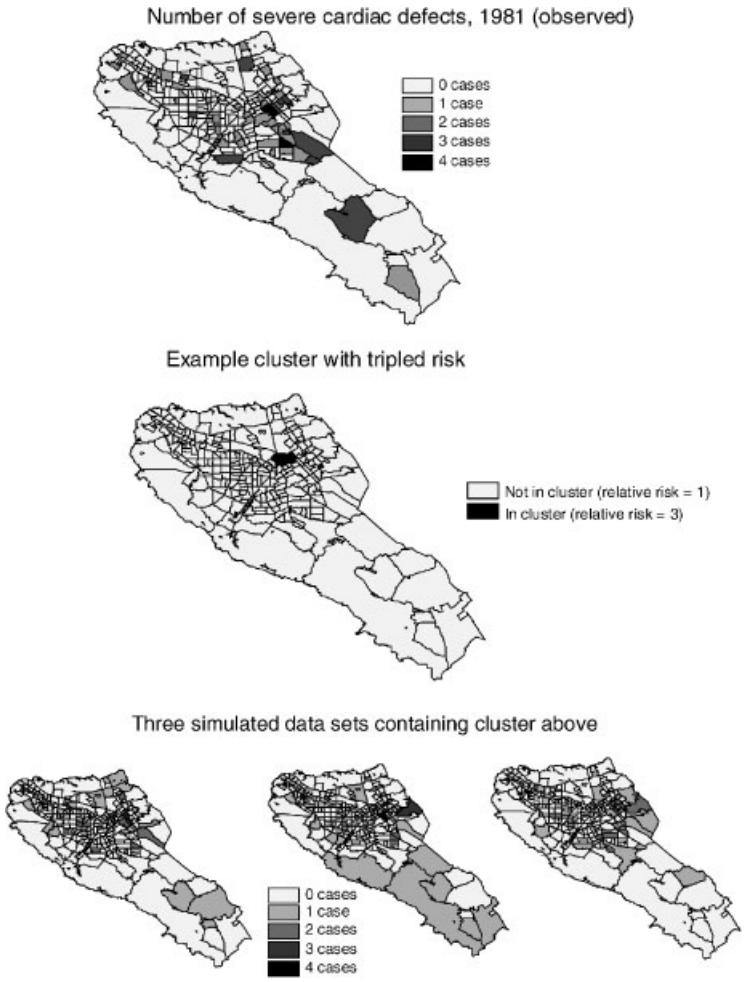

Figure 2. Map of the observed numbers of severe cardiac defects among reported live births in 1981 (top), an example 'cluster' comprised of a single tract and its six nearest neighbours (middle row), and three simulated data sets based on a tripling of disease risk in the indicated cluster (bottom row).

size. For larger tracts in the south or near the edges of the map, larger radii are needed to 'capture' the cluster of 7 neighbouring tracts. Also note that as τ increases, the highest observed power value decreases. This is most likely due to higher spatial weights assigned to tracts outside the smaller clusters, effectively diluting the observed impact of the generated clusters on Tango's statistic.

Figure 4 maps power quintiles for Kulldorff's [11] spatial scan statistic, with a maximum power of 36.5 per cent where we estimate power as the proportion of times the null hypothesis is rejected and the centre of the most likely cluster is among the 7 tracts defining the 'true' underlying cluster (middle map, right-hand side). Other related performance measures also appear in Figure 4. In general, patterns mirror a smoothed version of the geographic

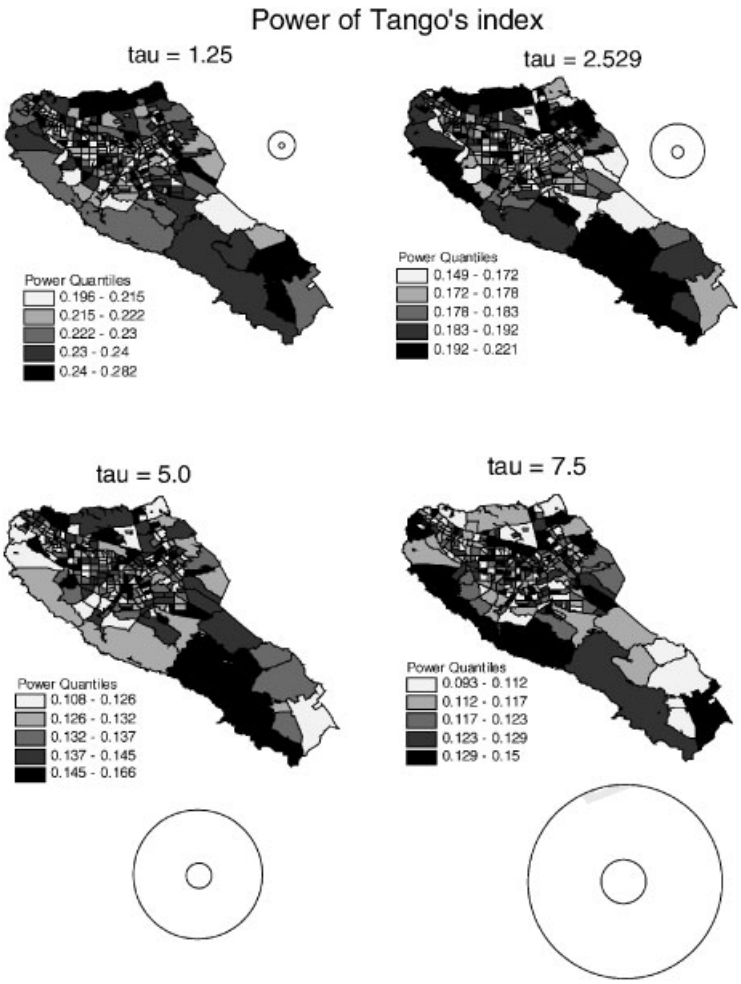

Figure 3. Choropleth maps of the statistical power of Tango's [8] test of clustering for various values of τ. Shading within each map is according to the quantiles of power values. Circles represent distance radii associated with neighbourhood weights of 0.5 (inner circle) and 0.05 (outer circle).

distribution of the population at risk (bottom map, Figure 1), with higher performance in areas with more reported live births. The lowest power values (near the 5 per cent level of the test) occur in the northwestern corner of the county. Comparing the middle and bottom row of maps contrasts local statistical performance (middle row) with 'location accuracy' (bottom row). Differences in these two types of performance include reduced probabilities of detection in areas with low numbers of live births such as tracts along the central western and northern borders.

The examples illustrate that adequate power summaries of tests to detect clusters and evaluate clustering depend on local characteristics of the geographic region under study, particularly the location of any real or suspected cluster or clusters.

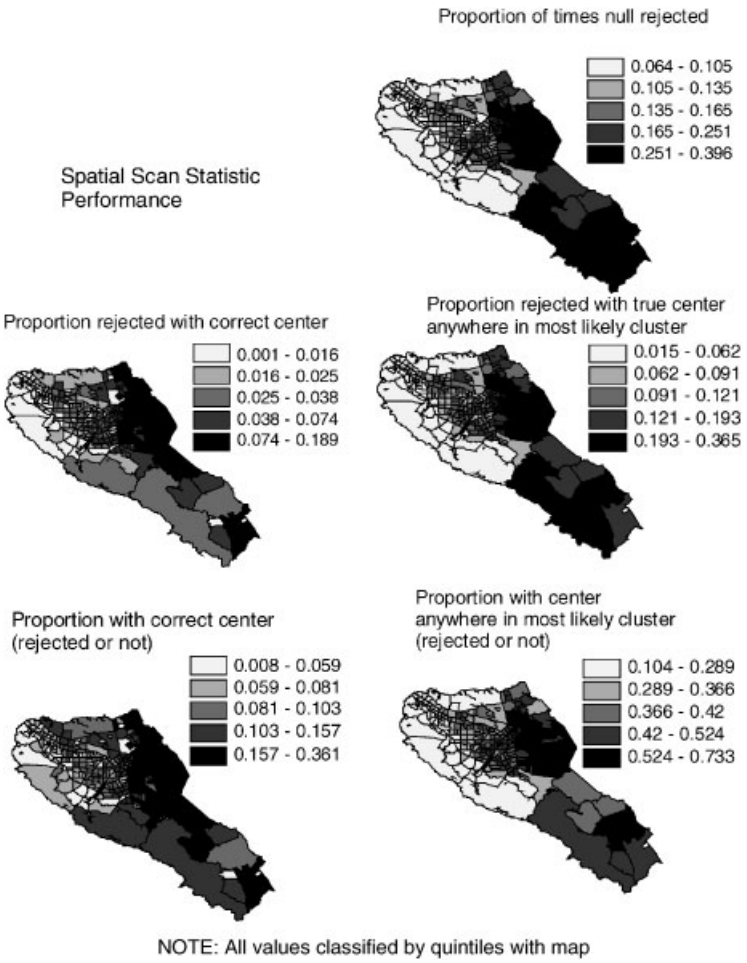

Figure 4. Choropleth map of the statistical power of Kulldorff's spatial scan statistic. Each tract is shaded according to the power of the test to detect clustering induced by a single cluster, centred at that particular tract, defined by tripling the relative risk of severe cardiac birth defects within the tract and its six nearest neighbouring tracts (see text).

4. DISCUSSION

In traditional power analysis, the two primary factors influencing power are sample size and effect size (the true difference in parameters between the null and alternative hypotheses). Not surprisingly, we observe the impact of the same two factors here, but with a geographic dimension. By comparing the performance of the tests in Figures 3 and 4 to the census tract populations at risk in Figure 1, we observe the power of the tests depends upon the size of the local population at risk (i.e. *local* sample size) for the tracts near the simulated disease cluster, however the local sample size does not provide the entire story. The examples relating to Tango's statistic in Figure 3, suggest that power is also dependent on the scale

parameter weighting collections of tracts. Pinpointing the separate and interacting impacts of local sample size and test parameters merits further investigation.

Kulldorff *et al.* [16] provide related discussion on the interplay between local effect sizes and local sample size. While we choose a constant relative risk and explore variations in power, Kulldorff *et al.* [16] consider varying local relative risk (local effect sizes) to equalize local power (at a 99.9 per cent level) in their discussion of power comparisons for disease clustering tests and development of benchmark data sets for comparison of tests. The goal of benchmark data is to provide a common test bed for detection methods which allows comparison to any past tests evaluated on the same benchmark data. Removing local power variations (through local variations in the effect size) makes sense in this setting, allowing one to calculate a single, overall power value for comparing tests with those evaluated on the same data in the past. It is worth noting that some local effect size variations reported in Kulldorff *et al.* [16] are quite extreme (ranging from relative risks within clusters of approximately 1.5 in some highly populated urban counties to over 190.0 in some sparsely populated rural counties). Hence, we feel there is also value in measuring and mapping variations in power associated with a fixed relative risk increase at a variety of locations, reflecting, for instance, variations in disease risk associated with an identical release of an environmental contaminant at a variety of different locations.

In hindsight, the conclusion that the power of a test to detect geographically local departures from the null hypothesis depends on the local sample size may seem natural (or even obvious). However, we find the maps of power informative (and often sobering), indicating areas in which clusters of a given effect size are more likely to be detected than others. Noting the distribution of disease cases under the null and alternative hypotheses of (2) and (5), we observe that in a study region of over 20 000 births, a local three-fold increase in the incidence rate of a very rare outcome (severe cardiac defects) in tracts with few births would be difficult to assess visually in Figure 2 or to detect using either of the tests considered. In census tracts with more births, detecting such an increase in cases is more likely.

While others report power studies based on clusters in varying locations with a study area [9, 16, 17] we are not aware of any previously published maps of power such as those shown here. We believe the example effectively illustrates the impact of geographic heterogeneity (in this case, of the population density) and test parameters on the statistical performance of hypothesis tests. In particular, the power of any test to detect departures from the null hypothesis can vary depending on where such departures might occur within the study area. Thus comparisons between tests of spatial pattern in heterogeneous populations depend not only on the strength and type of clustering [17, 18], but also depend on the locations where clusters occur within the data set at hand.

This analysis encompasses a '*post hoc*' theoretical power analysis as discussed by Waller and Poquette [17] who note the difficulty in interpretation of significant or non-significant statistical test results in a specific data setting without some notion of the power of the test over the study area. This is most revealing for a test with low power, where non-significant results allow a relatively large probability of failure to reject the null hypothesis when the specified alternative is true. Perhaps the phrase 'conditional power analysis' is more accurate than '*post hoc* power analysis' since the power analyses by Waller and Poquette [17] and above are conditional on the distribution of the population at risk but not on the observed distribution of disease cases. That is, the simulations above provide Monte Carlo estimates of the theoretical power of tests to detect the simulated clusters, and differ from the truly

post hoc notion of 'observed power' which is entirely dependent on the observed *p*-value of the test for the observed data [19].

The results above also illustrate that there is unlikely to be a single, omnibus test covering all alternatives of interest and performing equally well for clusters in any location. We intend the analyses above as an illustration of the impact of the location of a hypothesized cluster on statistical power and not as a comprehensive comparison of the tests considered. In fact, one could argue that the simulations above are poor assessments of performance of Tango's [12] test of *clustering* since we simulate only a single cluster. One often conceptualizes 'clustering' as a feature of the disease process acting over all cases, rather than a local anomalous region of increased risk better detected by tests to detect particular clusters. To be fair, the reduced power of Tango's [12] statistic in our example is partially due to the fact that the majority of the 259 census tracts do follow the null hypothesis, only 7 of these experience increased relative risk.

In summary, the example above illustrates the dependence between the statistical power of tests of disease clustering and the strength, type, and location of suspected disease clusters. The example also suggests the observed spatial distribution of the population at risk often provides a necessary context for interpreting power comparisons between different methods of assessing the spatial patterns of disease.

ACKNOWLEDGEMENTS

The first two authors were partially supported by Associations of Schools of Public Health/Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry Cooperative Agreement Project number S779-18/18, and the first author by National Institute of Environmental Health Sciences Grant R01-ES007750. The opinions expressed above represent those of the authors and do not necessarily reflect those of the funding agencies. The research was conducted in part while the third author was a graduate student in the Department of Biostatistics, Rollins School of Public Health, Emory University. The authors thank two anonymous referees for constructive comments on an earlier version of this manuscript.

REFERENCES

1. Elliott P, Martuzzi M, Shaddick G. Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research* 1995; **4**:137–159.
2. Tango T. Comparison of general tests for spatial clustering. In *Disease Mapping and Risk Assessment for Public Health*, Lawson A, Biggeri A, Böhning D, LeSaffre E, Viel J, Bertollini R (eds). Wiley: Chichester, 1999.
3. Wakefield JC, Kelsall JE, Morris SE. Clustering, cluster detection, and spatial variation in risk. In *Spatial Epidemiology*: *Methods and Applications*, Elliott P, Wakefield JC, Best NG, Briggs DJ (eds). Oxford University Press: Oxford, 2000.
4. Kulldorff M. Statistical methods for spatial epidemiology: Tests for randomness. In *GIS and Health in Europe*, Loytonen M, Gatrell A (eds). Taylor & Francis: London, 1998.
5. Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*. Wiley: Hoboken, NJ, 2004.
6. Besag J, Newell J. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society*, *Series A* 1991; **154**:143–155.
7. Kulldorff M, Nagarwalla N. Spatial disease clusters-detection and inference. *Statistics in Medicine* 1995; **14**:799–810.
8. Hill EG, Ding L, Waller LA. A comparison of three tests to detect general clustering of a rare disease in Santa Clara County, California. *Statistics in Medicine* 2000; **19**:1363–1378.
9. Gangnon RE, Clayton MK. A weighted average likelihood ratio tests for spatial clustering of disease. *Statistics in Medicine* 2001; **20**:2977–2987.
10. Shaw G, Selvin S, Swan S, Merrill D, Schulman J. An examination of three spatial disease clustering methodologies. *International Journal of Epidemiology* 1988; **17**:913–919.

11. Kulldorff M. A spatial scan statistic. *Communications in Statistics*: *Theory and Methods* 1997; **26**:1487–1496.
12. Tango T. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine* 1995; **14**:2323–2334.
13. Waller LA, Jacquez GM. Disease models implicit in statistical tests of disease clustering. *Epidemiology* 1995; **6**:584–590.
14. Swan S, Shaw G, Harris J, Neutra R. Congenital cardiac anomalies in relation to water contamination, Santa Clara County, California, 1981–1983. *American Journal of Epidemiology* 1989; **129**:885–893.
15. Kulldorff M, Information Management Services, Inc. *SaTScan*™ *v 5.1*: *Software for the Spatial and Space-time Scan Statistics*, 2004. http://www.satscan.org/
16. Kulldorff M, Tango T, Park PJ. Power comparisons for disease clustering tests. *Computational Statistics and Data Analysis* 2003; **42**:665–684.
17. Waller LA, Poquette CA. The power of focused score tests under misspecified cluster models. In *Disease Mapping and Risk Assessment for Public Health*, Lawson A, Biggeri A, Böhning, D, LeSaffre E, Viel J, Bertollini R (eds). Wiley: Chichester, U.K., 1999.
18. Waller LA, Lawson AB. The power of focused tests to detect disease clustering. *Statistics in Medicine* 1995; **14**:2291–2308.
19. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* 2001; **55**:19–24.

