International Journal of Health Geographics

BioMed Central

Research

Open Access

Local clustering in breast, lung and colorectal cancer in Long Island, New York

Geoffrey M Jacquez[1,2] and Dunrie A Greiling*[1,2]

Address: [1]TerraSeer, Inc., Ann Arbor, MI, USA and [2]BioMedware, Inc., Ann Arbor, MI USA

Email: Geoffrey M Jacquez - jacquez@biomedware.com; Dunrie A Greiling* - dunrie@biomedware.com

* Corresponding author

Published: 17 February 2003

International Journal of Health Geographics 2003, **2**:3

Received: 10 February 2003

Accepted: 17 February 2003

This article is available from: http://www.ij-healthgeographics.com/content/2/1/3

Abstract

Background: Analyses of spatial disease patterns usually employ a univariate approach that uses one technique to identify disease clusters. Because different methods are sensitive to different aspects of spatial pattern, an approach employing a battery of techniques is expected to describe geographic variation in human health more fully. This two-part study employs a multi-method approach to elucidate geographic variation in cancer incidence in Long Island, New York, and to evaluate spatial association with air-borne toxics. This first paper uses the local Moran statistic to identify cancer hotspots and spatial outliers. We evaluated the geographic distributions of breast cancer in females and colorectal and lung cancer in males and females in Nassau, Queens, and Suffolk counties, New York, USA. We calculated standardized morbidity ratios (SMR values) from New York State Department of Health (NYSDOH) data.

Results: We identified significant local clusters of high and low SMR and significant spatial outliers for each cancer-gender combination. We then compared our results with the study conducted by NYSDOH using Kulldorff's spatial scan statistic. We identified patterns on a smaller spatial scale with different cluster shapes than the NYSDOH analysis did, a consequence of different statistical methods and analysis scale.

Conclusion: This is a methodological and comparative study to evaluate whether there is substantial benefit added by using a variety of techniques for geographic pattern detection at different spatial scales. We located significant spatial pattern in cancer morbidity in Nassau, Queens, and Suffolk counties. These results broadly agree with the results of other studies that used different techniques, but differ in specifics. The differences in our results and that of the NYSDOH underscore the need for an exploratory, integrative, and multi-scalar approach to assessing geographic patterns of disease, as different methods identify different patterns. We recommend that future studies of geographic patterns use a concordance of evidence from a multiscalar integrative geographic approach to assure that 1) different aspects of spatial pattern are fully identified and 2) the results from the suite of analyses are logically consistent.

Background

This paper is the first in a two-paper series. It addresses whether there are statistically significant clusters of cancer on Long Island. The second paper [1] examines whether the patterns in cancer are spatially associated with patterns in the environment, specifically air toxics. This set of two studies were prompted by ongoing concern over cancer patterns on Long Island. Citizens and public health

workers are concerned about two things – whether cancer clusters exist and, if so, what may explain the clusters. The possibility of breast cancer clusters on Long Island has been in the news and the focus of recent research [2]. New York state had the 4th highest death rate from breast cancer in 1995–99, though it was 17th in colorectal cancer and 39th in lung cancer [3]. While breast cancer rates are higher in the Northeastern US than in other parts of the country, Kulldorff et al. [2] established that the entire New York-Philadelphia metropolitan area has higher breast cancer mortality rates than the remainder of the Northeast. The analysis performed by the New York State Department of Health (which used the average cancer incidence/population for New York as a whole as a reference) located significant elevations in breast cancer on Long Island in particular. Even when compared to New York state as a whole, the cancer rates on Long Island seem to be elevated. Given the concern and the apparent elevation of cancer rates on Long Island, we focused our study on Nassau, Queens, and Suffolk counties, the easternmost three counties on Long Island.

Methods

Data

The New York State Department of Health (NYSDOH) published the cancer incidence data online as part of their Cancer Surveillance Improvement Initiative, http://www.health.state.ny.us/nysdoh/cancer/csii/nyscsii.htm. These data represent newly diagnosed cancer cases in the period 1993–7 assigned to the patient's residence at diagnosis, and they are calculated as the number of cancers for each 100,000 people in the population. When we began this study (August 2001), the NYSDOH had released data on three cancers: breast (female only), colorectal (female and male), and lung (female and male) cancers. Since then, they released data on prostate cancer for the years 1994–8, which we did not include in this study. Data has not yet been added for years other than 1993–7 for the three cancers we analyzed.

To protect patient privacy, the NYSDOH data provided case counts referenced to ZIP codes rather than individual residences. ZIP codes are regions developed for mail delivery by the US Postal Service. In the study area, the population in ZIP codes ranges between 445–105,723 individuals, with a mean of about 23,000 (using 2000 US Census numbers, http://factfinder.census.gov). They are not uniform in population nor ethnicity nor age. While ZIP codes are somewhat arbitrary spatial units of analysis with respect to potential health and environmental factors, they provided NYSDOH a convenient way to group the population to protect patient confidentiality. Data at a better spatial resolution were not made available to us. We combined the cancer diagnosis data with ZIP code boundary files, reflecting the geography in November 1999. We purchased the boundary files from Claritas Corporation http://www.claritas.com. While the NYSDOH provides information on the entire state, we focus on the 214 ZIP codes within Nassau, Queens and Suffolk County on Long Island.

People move between ZIP codes and cancer latency (the time between causative exposures and cancer onset) is estimated to be between 5–40 years for these cancers, so the ZIP code where the patient was diagnosed may not be the location where the cancer developed nor where causative exposures occurred. We do not include any adjustments for migration or changes in any demographic patterns within the study area.

While the observed cancer diagnosis data did adjust for different populations-at-risk in the different ZIP codes, we also used New York State's adjustment for different age patterns as well. Because cancer incidence is related to age, NYSDOH calculated the expected cancer incidence for each ZIP code using the ZIP code's age structure and the average incidence by age class for New York state. We calculated a standardized morbidity ratio (SMR) by dividing the observed value by the age-adjusted expected incidence. An SMR value of 1.0 indicates that the observed incidence is the same as expected, lower than 1.0 indicates that fewer than expected cases of cancer occurred, and greater than 1.0 indicates that more than expected occurred.

Local Cluster Analysis

We identified significant clustering and spatial outliers in SMR using Anselin's local Moran test [6] in the ClusterSeer™ software http://www.terraseer.com/clusterseer.html. The local Moran test evaluates local clustering or spatial autocorrelation by evaluating the contribution of each location to the Moran's I statistic for the whole study area. Its null hypothesis is that there is no association between SMR values in neighboring ZIP codes (no spatial autocorrelation). The working (alternative) hypothesis is that spatial clustering exists. The statistic is:

$$I_i = Z_i \sum_j W_{ij} Z_j$$

where I_i is the statistic for ZIP code i, z_i is the difference between the SMR at i and the mean SMR for Long Island, z_j is the difference between the SMR at j and the mean for Long Island. w_{ij} is a weight so that the statistic only considers neighbors that share a common border (w_{ij} is 1/n if the two ZIP codes are neighbors using the rook contiguity relationship, where n is the number of rook neighbors, and w_{ij} is zero otherwise). We evaluated the test statistic using Monte Carlo P-values [6], obtained from 99,999

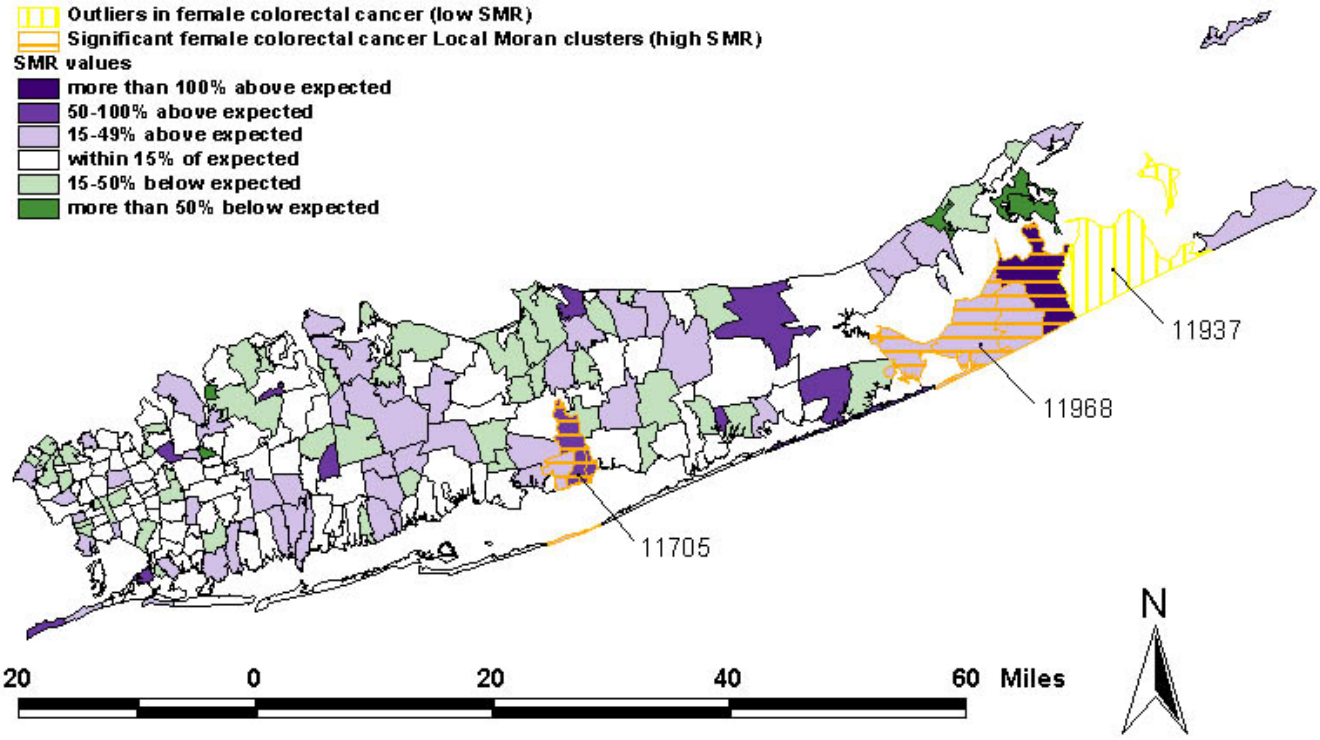

Figure 1

Geographic distribution of female colorectal cancer. The fill color in each ZIP code represents the SMR, with green indicating relatively low SMR and purple representing relatively high SMR. White indicates SMR near 1 (observed and expected equivalent). The ZIP codes outlined and cross-hatched in orange had significantly high incidences and formed local clusters under Moran's test. The ZIP code outlined and cross-hatched in yellow was a significant spatial outlier by the local Moran test, though its SMR is not significantly different from 1. The black outlines describe ZIP code boundaries. Labels identify the centering ZIP codes for each cluster or outlier. The strip of Fire Island outlined in orange is part of ZIP code 11782, the main portion of which is a neighbor of 11705 on Long Island proper.

conditional randomizations of the dataset. Because it is possible for local clusters to overlap because of shared neighbors, we therefore used an adjusted significance level (Bonferroni adjustment) to evaluate P-values ($\alpha' = 0.01101$).

The local Moran test as we used it evaluates ratio data, in this case the SMR. It does not, however, consider whether that ratio is based on many or few cases. A ratio based on few observed or expected cases (small ZIP code population) is more unstable than one based on more cases. To evaluate the stability of the SMR for significant clusters, we calculated its confidence interval [7]. The confidence intervals increase as the observed number of cases decreases. Confidence intervals that overlap 1.0 indicate no statistically significant difference from the expected ratio.

Results

Colorectal Cancer-Local Cluster Analysis

Females

The local Moran test identified two local clusters with SMRs about 45–50% higher than the New York average, centered on Bayport (ZIP 11705) and Southampton (11968) (Table 1, Figure 1). Wainscott (11937) is a significant spatial outlier with an SMR 70% of New York state's average. Wainscott's SMR, however, has wide confidence intervals due to the small number of observed cases there. Thus, while statistically distinct from its neighbors, it does not have significantly reduced risk.

Males

For colorectal cancer in males, the local Moran test identified five local clusters with SMR values over 50% higher than the New York average (Table 2, Figure 2). These clusters share some ZIP codes in common; they form two large rather than five small clusters. The western cluster is

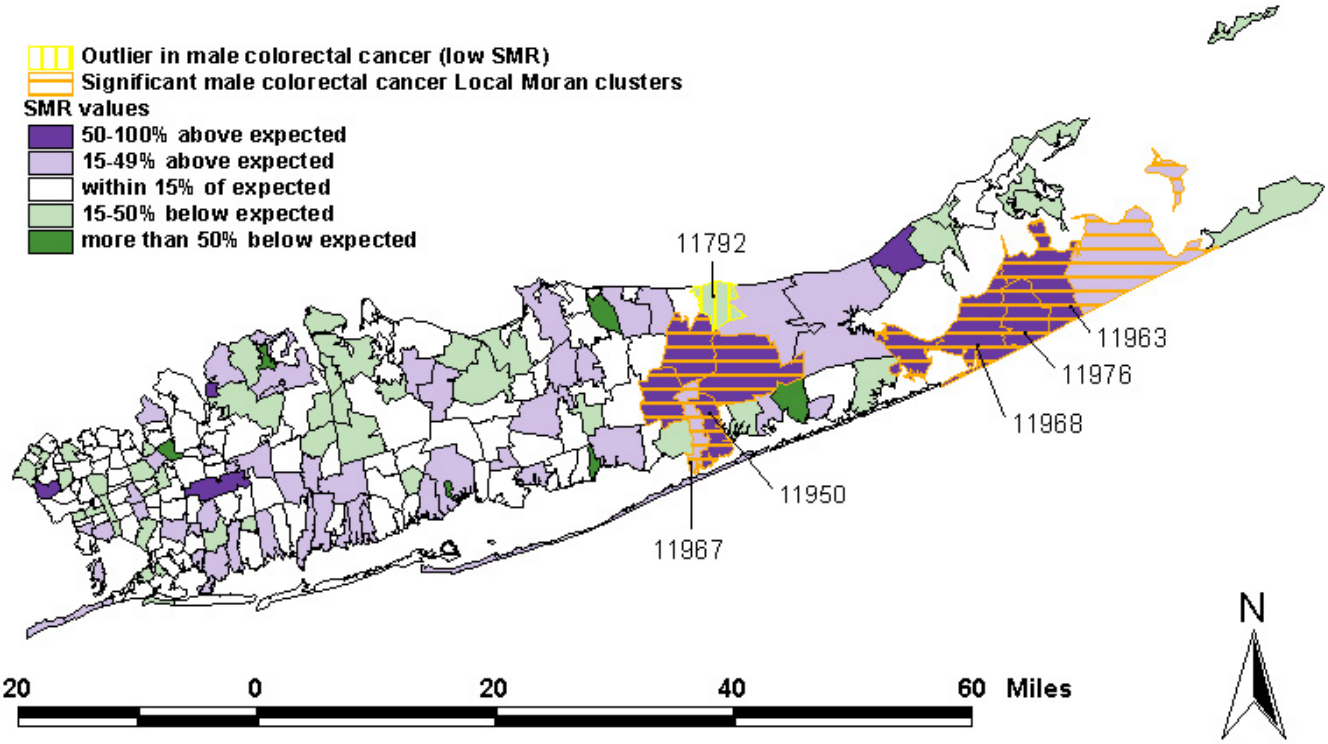

Figure 2
Geographic distribution of male colorectal cancer. The fill color in each ZIP code represents the SMR, with green indicating relatively low SMR and purple representing relatively high SMR. White ZIP codes indicate SMR near 1 (observed and expected equivalent). The ZIP codes outlined and cross-hatched in orange had significantly high incidences and formed local clusters under Moran's test. The ZIP code outlined and cross-hatched in yellow is a significant spatial outlier in the local Moran analysis, though its confidence interval is not significantly different from 1. The black outlines describe ZIP code boundaries. Labels identify the centering ZIP codes for each cluster or outlier.

comprised of two significant local clusters centered on Shirley (11967) and Mastic (11950). The eastern cluster has three significant local clusters centered on Southampton (11968), Water Mill (11976), and Sagaponak (11963). Wading River (11792) is a significant spatial outlier with an SMR about 70% of the New York state average. While Wading River is a significant outlier under the local Moran test, its SMR has a wide confidence interval due to the small number of observed cases there. Thus, while statistically distinct from its neighbors, it does not have significantly reduced risk.

Breast Cancer – Local Cluster Analysis

We identified two local clusters with SMR's 70–83% of the New York average. These significant local clusters of low cancer overlap to form one larger cluster, centered on Floral Park (ZIP 11103) and Woodside in Flushing (11137) (Table 3, Figure 3). The local Moran test also detected two local clusters with SMR 30–50% higher than the New York average. These significant local clusters also overlap and are centered on Southampton (11968) and Wainscott (11937). Shelter Island (11964) is a significant spatial outlier, though its confidence intervals are wide (and overlap 1) due to the small number of observed cases there. Thus, while statistically distinct from its neighbors, it does not have significantly reduced risk.

Lung Cancer – Local Cluster Analysis

Females

The local Moran test identified three clusters with SMR about 70% of the New York average (Figure 4, Table 4). These three local clusters are contiguous, share ZIP codes, and together comprise a single, large cluster extending through portions of Flushing in the north and Jamaica in the south. This cluster is centered on Flushing (ZIPs 11368 and 11367) and Saint Albans in Jamaica (11412). Sayville (11782) is a significant spatial outlier with low SMR (72% of the New York average), though its SMR has

Table 1: Colorectal cancer in females

ZIP Codes		Incidence Observed (O) per 100,000 population	Incidence Expected (E) per 100,000 population, adjusted for age using the NY State average	SMR (O/E)	95% Confidence Interval for SMR	Local Moran Statistic	Two-Tailed P-value
Centering Region	Included Neighbors						
11705	11782, 11741, 11715	94	64.6	1.4551	1.1888, 1.7811	3.1591	0.00745
11937	None	21	24.0	0.875	0.5705, 1.3402	-1.4860	0.00157
11968	11946, 11963, 11976	97	61.2	1.5850	1.2989, 1.9340	2.4271	0.00329

Table 2: Colorectal cancer in males.

ZIP Codes		Incidence Observed (O) per 100,000 population	Incidence Expected (E) per 100,000 population, adjusted for age using the NY State average	SMR (O/E)	95% Confidence Interval for SMR	Local Moran Statistic	Two-Tailed P-value
Centering Region	Included Neighbors						
11792	None	7	10.3	0.6796	0.3240, 1.4256	-1.6215	0.00553
11950	11951, 11967, 11949, 11955	93	58.5	1.5897	1.2974, 1.9480	3.5058	0.00109
11963	11976, 11968, 11937	109	68.7	1.5866	1.3150, 1.9143	4.1484	0.00006
11967	11719, 11980, 11961, 11949, 11950, 11951	148	94.5	1.5661	1.3331, 1.8399	2.3135	0.00004
11968	11946, 11963, 11976	105	62	1.6935	1.3987, 2.0505	6.5445	0.00006
11976	11968, 11963	73	43.1	1.6937	1.3465, 2.1305	7.3113	0.00149

a wide confidence interval resulting from the small number of observed cases there. Thus, while statistically distinct from its neighbors, it does not have significantly reduced risk.

Seven local clusters of high SMR values occurred in the more central portions of Long Island (Table 4, Figure 4). There is a cluster in north-mid Long Island, made up of two significant local clusters centered on Bayville (11709) and Mill Neck (11765). This cluster has about 60–70% higher SMR than the New York state average. A large cluster in south central Long Island is composed of four local clusters centered on Ronkonkama (11779), Central Islip (11722), Islip Terrace (11752), and East Islip (11730). This cluster has an SMR about 40% higher than the New York state average. Further east is a third cluster of high female lung cancer incidence centered on Mastic (11950) and including several adjacent ZIP codes. Its SMR is about 60% higher than the New York state average.

Males

Three local clusters of low SMR values were identified, centering on Great Neck (ZIP 11024), Roslyn (11576), and Huntington (11743), all in the northwest portion of Long Island (Figure 5, Table 5). These clusters are typified by lung cancer SMR values that are 50–75% of the New York State average. The two clusters centered on 11024 and 11576 are adjacent to one another forming a single, large cluster of low lung cancer SMR for males. In addition, Far Rockaway (11694) and Moriches (11955) are significant spatial outliers with 88% and 84% of the New York SMR respectively, though the SMR confidence inter-

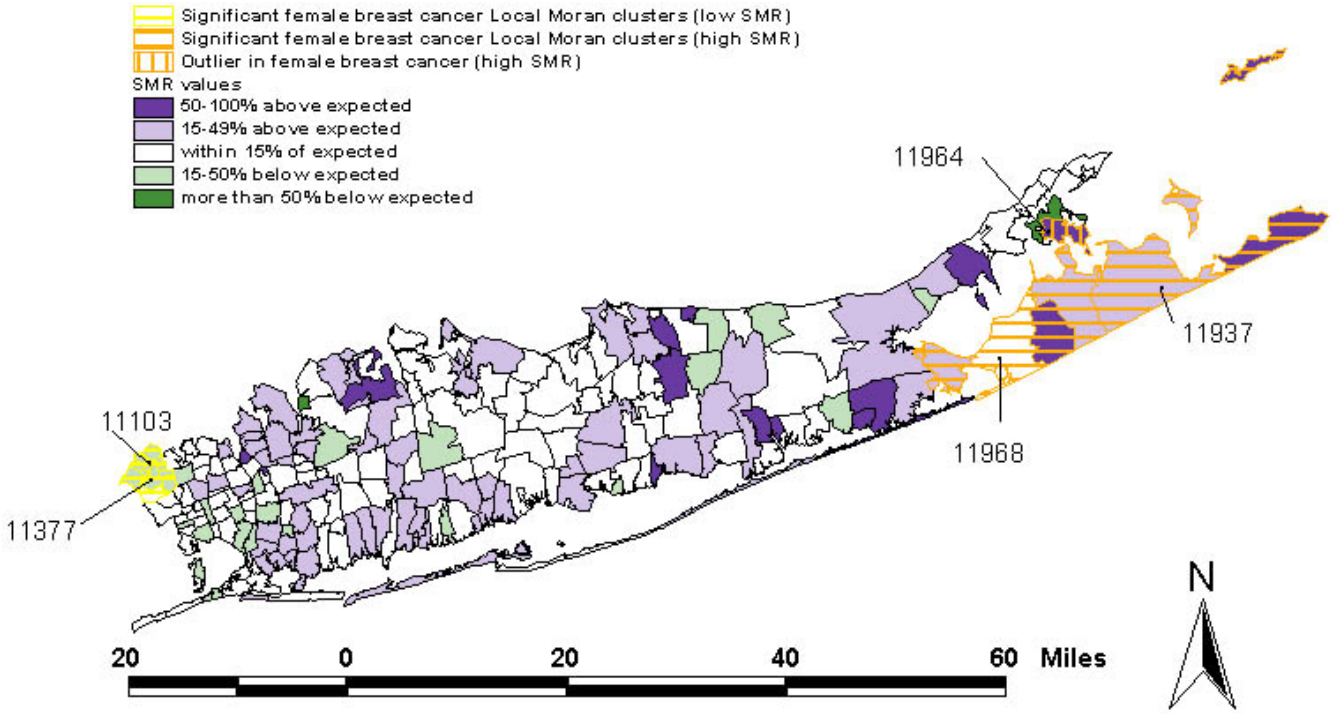

Figure 3

Geographic distribution of female breast cancer. The fill color in each ZIP code represents the SMR, with green indicating relatively low SMR and purple represent relatively high SMR. White ZIP codes indicate SMR near 1 (observed and expected equivalent). Orange and yellow hatching of ZIP codes indicate significant clusters and outliers according to the local Moran test. The black outlines describe ZIP code boundaries. Labels identify the centering ZIP codes for each significant cluster or outlier.

Table 3: Clusters of high and low breast cancer SMR in females.

ZIP Codes		Incidence Observed (O) per 100,000 population	Incidence Expected (E) per 100,000 population, adjusted for age using the NY State average	SMR (O/E)	95% Confidence Interval for SMR	Local Moran Statistic	Two-Tailed P-value
Centering Region	Included Neighbors						
11103	11101, 11106, 11102, 11105, 11370, 11377	718	953.3	0.7532	0.7000, 0.8103	2.1699	0.00165
11377	11378, 11104, 11101, 11103, 11370, 11372, 11373	1021	1305.9	0.7818	0.7353, 0.8313	1.3158	0.00345
11937	11963, 11954	138	97.9	1.4096	1.1930, 1.6655	2.0559	0.00937
11968	11946, 11963, 11976	159	120.7	1.3173	1.1277, 1.5388	0.1527	0.00801
11964	none	10	6.6	1.5152	0.8152, 2.816	-5.884	0.00479

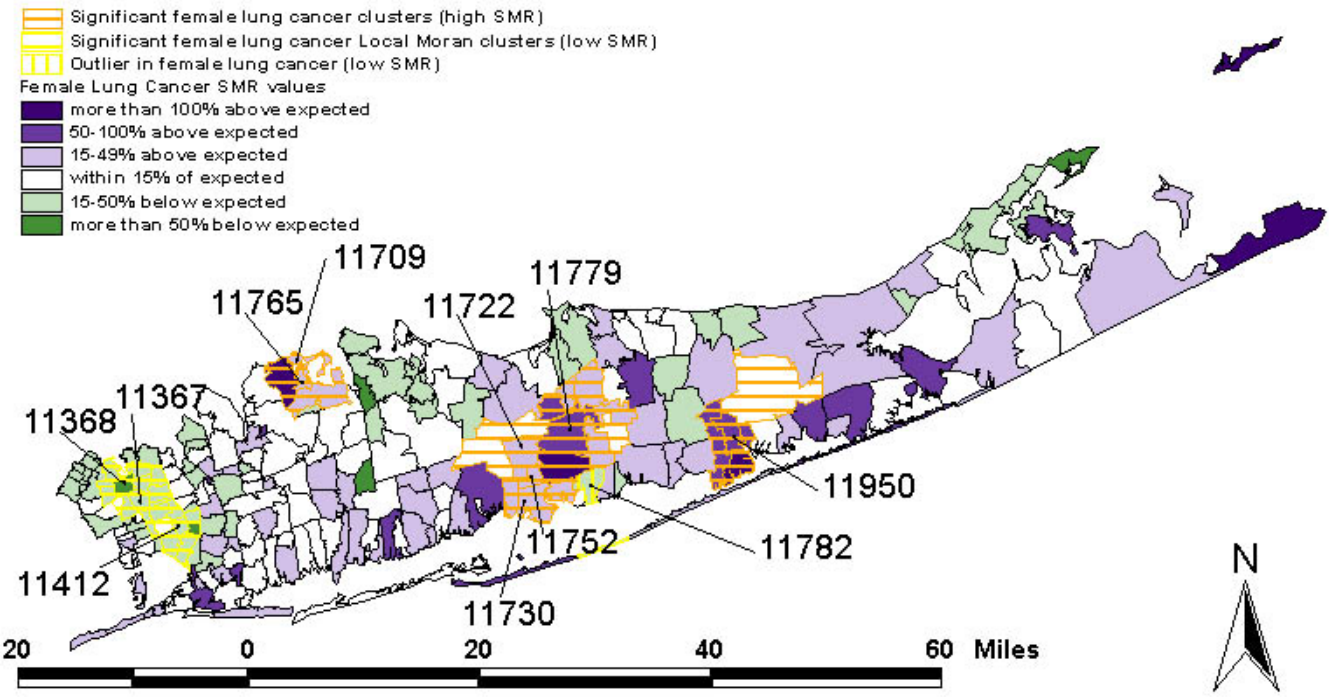

Figure 4

Geographic distribution of lung cancer in females. The black outlines describe ZIP code boundaries. The fill color in each ZIP code represents the SMR, with green indicating relatively low SMR and purple representing relatively high SMR. White ZIP codes indicate SMR near 1 (observed and expected equivalent). Orange and yellow hatching of ZIP codes indicate significant clusters and outliers according to the Local Moran test. Labels identify the centering ZIP codes for each significant cluster or outlier.

vals are wide because the number of observed cases is small in each location. Thus, while statistically distinct from its neighbors, it does not have significantly reduced risk.

A large cluster of lung cancer SMR 20–60% higher than the New York average was identified in central Long Island. This larger cluster is composed of 9 significant local clusters, centered on Farmingville (11738), Coram (11727), Miller Place (11764), Middle Island (11953), Mastic (11950), Mastic Beach (11951), Shirley (11967), Medford (11763), and Sayville (11782).

Discussion

Comparison to Prior Studies

In this section we compare our results to the New York state maps of cancer morbidity and compare and contrast them to the geographic variation patterns identified by the local Moran statistic and by boundary analysis [1]. New York State used Kulldorff's spatial scan statistic to evaluate the significance of geographic patterns in cancer http://www.health.state.ny.us/nysdoh/cancer/csii/nysc-sii.htm, [4,5].

Qualitative differences in clusters

Some differences are immediately apparent when one compares, for breast cancer, the scan statistic clusters http://www.health.state.ny.us/nysdoh/cancer/csii/nysc-sii.htm to the local Moran clusters (Figure 3). Under the scan statistic, all of eastern and most of western Suffolk are declared a cluster, as are substantial portions of Nassau, and the western portions of Long Island. In contrast, the local Moran statistic finds significant clustering only in the southern fork towards Montauk, and identifies a significant clustering of low breast cancer morbidity on western Long Island. Boundaries in breast cancer morbidity occur throughout Long Island, and identify adjacent ZIP codes that differ substantially in cancer morbidity [1]. Hence the local Moran and boundary approaches identify clusters on a finer spatial scale, while the scan approach is identifying larger clusters.

Table 4: Lung cancer in females.

ZIP Codes		Incidence Observed (O) per 100,000 population	Incidence Expected (E) per 100,000 population, adjusted for age using the NY State average	SMR (O/E)	95% Confidence Interval for SMR	Local Moran Statistic	Two-Tailed P-value
Centering Region	Included Neighbors						
11367	11375, 11368,11355, 11365, 11366, 11432,11435	449	639.4	0.7022	0.6402, 0.7703	0.7465	0.00715
11368	11373, 11372, 11369, 11354, 11355, 11367, 11375, 11374	603	862.3	0.6993	0.6456, 0.7574	1.5671	0.00864
11412	11434, 11433, 11423, 11429, 11411, 11413	219	324.4	0.6751	0.5913, 0.7707	0.8040	0.00453
11709	11560, 11771	61	37	1.6486	1.2827, 2.1189	0.8983	0.00631
11722	11717, 11788,11779, 11716,11752, 11751	271	195.3	1.3876	1.2319, 1.5631	1.1527	0.00504
11730	11751, 11752, 11716, 11769	118	76.2	1.5486	1.2929, 1.8548	1.3970	0.00332
11752	11751, 11722, 11716, 11730	149	97.2	1.5329	1.3055, 1.7999	1.4626	0.00228
11765	11560, 11771	51	28.7	1.7770	1.3505, 2.3382	0.7465	0.006310
11779	11722, 11788, 11767, 11755, 11720, 11738, 11742, 11741, 11716	310	218.1	1.4214	1.2716, 1.5887	1.5155	0.00843
11782	None	17	23.4	0.7264	0.4516, 1.1687	-1.1525	0.00246
11950	11967, 11949, 11955, 11951	90	54.8	1.6423	1.3358, 2.0192	3.7441	0.00094

In addition, the scan statistic clusters include groups of ZIP codes that, in fact, have cancer morbidity *below* the New York State average. For example, the North Fork is deemed part of a significant cluster of elevated incidence, even though the contiguous ZIP codes 11971 (SMR = 0.9058) and 11944 (SMR = 0.8602) are below the expected value. Why might this be? First, the scan statistic evaluates geographic relationships using centroids and a circular scan window. Hence relevant geographies, such as Peconic bay, are not taken into account. In effect, the cancer morbidities on the North Fork are lumped with the cancer morbidities on the south fork. The local Moran approach, by evaluating geographic relationships using common borders, doesn't connect ZIP codes on the North Fork to ZIP codes on the South Fork. It is indeed a "local" statistic (hence its name) and is sensitive only to local clusters of cancer morbidity. Second, the scale of the study is different for the scan statistic and for the local Moran and boundary analyses. The NYS Department of Health applied the scan method to New York as a whole, while the local Moran and boundary analyses used only the ZIP codes in Long Island. Each of these techniques employs randomization methods to evaluate statistical significance, and the results therefore may vary when one includes more areas in the analysis.

To summarize, the local Moran clusters and the boundaries in breast cancer incidence quantify spatial pattern on a finer spatial scale than scan statistic clusters. This arises because the local Moran and boundary analysis techniques use ZIP code zone adjacency (sharing of a common border) rather than distance between centroids to evaluate geographic relationships, and because the scale of the study differs.

The scale of the process

This general result – the local Moran results and the boundary analysis results identify finer scale variation, while the scan approach picks up larger scale clusters – holds for all cancers considered and is to be expected giv-

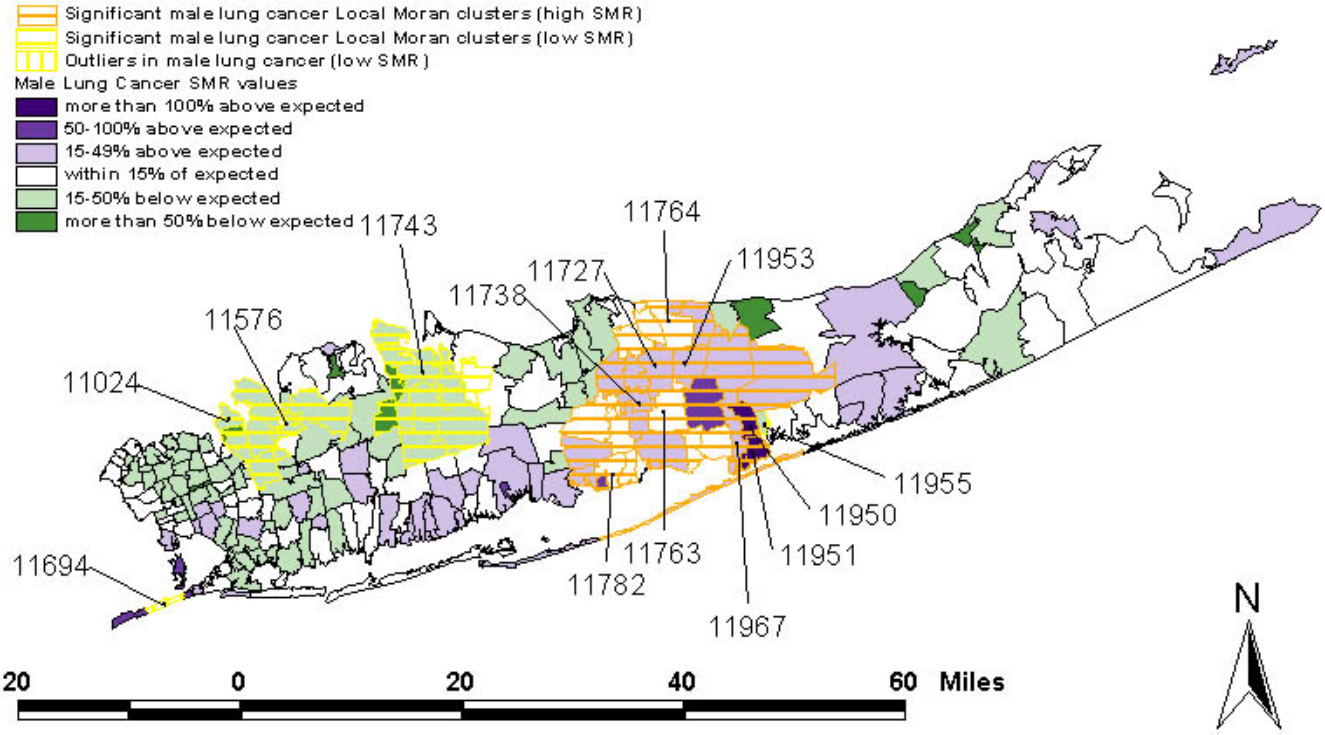

Figure 5

Lung cancer in males. The fill color in each ZIP code represents the SMR, with green indicating relatively low SMR and purple representing relatively high SMR. White ZIP codes indicate SMR near 1 (observed and expected equivalent). Orange and yellow hatching of ZIP codes indicate significant clusters and outliers according to the Local Moran test. The blue outlines describe ZIP code boundaries. Labels identify the centering ZIP codes

en the differences in the methods. An understanding of how results vary as a function of the scale of the study and as a function of the spatial sensitivity of the method is important for us to evaluate geographic variation in cancer morbidity. We should expect potential exposures as well as cancer correlates and covariates to have local- as well as large-scale components. At the local scale neighborhoods can change dramatically from one location to another, and many putative exposures involve point sources that have local impacts (for example, fuel leaks from underground storage tanks are often limited to a few hundred cubic yards of contaminated soil). But we also know that wind-borne pollutants can impact much larger areas, as can agricultural pesticides and contaminants carried in ground water. Hence the techniques we employ should be sensitive to multi-scale geographic heterogeneity, and thus able to identify the fine local scale variation that may be underlying exposures and/or genetic differences at the local level, as well as regional and sub-regional patterns.

The scale of the study

While both the local Moran and scan statistics assess statistical significance of a cluster relative to the morbidity for New York state as a whole, we consider only Long Island ZIP codes in the local Moran analysis, while the NYS scan analysis considers ZIP codes for all of New York state. Both statistics (local Moran and scan) use randomization techniques that "sprinkle" morbidity values over the study area in order to construct reference distributions under the null hypothesis. Thus the null hypothesis for the local Moran analysis is no spatial structure in breast cancer incidence in Long Island, whereas the null hypothesis for the scan statistic is no spatial structure in breast cancer incidence in all of New York State. Our randomization approach "resampled" the data under the null hypothesis that a cancer incidence observed in a Long Island ZIP code is equally likely to occur in any other Long Island ZIP code. This means cancer risk on was assumed to be equal across Long Island. By randomizing across all of New York State, the NYSDOH study assumed cancer risk to be equal across New York State and that, for example, risk on Long Island is the same as risk in the Adirondacks. According to

Table 5: Lung cancer in males

ZIP Codes		Incidence Observed (O) per 100,000 population	Incidence Expected (E) per 100,000 population, adjusted for age using the NY State average	SMR (O/E)	95% Confidence Interval for SMR	Local Moran Statistic	Two-Tailed P-value
Centering Region	Included Neighbors						
11024	11023, 11021	42	80.7	0.5204	0.3846, 0.7042	1.3050	0.01090
11576	11020, 11030, 11050, 11545, 11548, 11568, 11577, 11507, 11596, 11040	255	386.5	0.6598	0.5836, 0.7459	1.3510	0.00255
11694	None	45	51	0.8824	0.6588, 1.1818	-0.41257	0.00372
11727	11738, 11784, 11776, 11766, 11764, 11953, 11763	247	197.1	1.2532	1.1062, 1.4196	1.6698	0.00755
11738	11779, 11720 11784, 11727 11763, 11742	274	222.8	1.2298	1.0925, 1.3844	1.6413	0.00675
11743	11797, 11724, 11721, 11740, 11731, 11746, 11747	259	357.8	0.7239	0.6409, 0.8176	1.0827	0.00535
11763	11742, 11738, 11727, 11953, 11980, 11713, 11772	260	203.3	1.2789	1.1325, 1.4442	0.7947	0.00051
11764	11727, 11766, 11789, 11778, 11953	137	108.3	1.2650	1.0700, 1.4956	0.2423	0.00874
11782	11706, 11796, 11769, 11716, 11741, 11705, 11772	323	269	1.2007	1.0767, 1.3391	0.5181	0.00719
11950	11967, 11949, 11955, 11951	116	78	1.4872	1.2397, 1.7840	5.6718	0.00348
11951	11967, 11950	86	53.6	1.6045	1.2988, 1.9821	8.9239	0.0022
11953	11763, 11727, 11764, 11778, 11961, 11980	204	161.1	1.2663	1.1039, 1.4526	1.8116	0.01048
11955	None	5	5.9	0.8474	0.3528, 2.0361	-0.55418	0.00240
11967	11719, 11980, 11961, 11949, 11950, 11951	178	123.4	1.4425	1.2454, 1.6707	1.6379	0.00001

the NYSDOH study, much of Long Island is included in clusters defined by higher incidence of breast cancer, male and female colorectal cancer, and female lung cancer.

This illustrates the spatial scale of the study is closely related to the question(s) being addressed. For the local Moran analysis, we are asking questions specific to Long Island, against a null hypothesis that states that cancer risk is uniform across Long Island. In their analysis using the scan statistic, the NYSDOH is addressing questions regarding cancer incidence in all of New York State, against the implicit assumption that risk is uniform across the entire state. However, the standardized morbidity ratio (SMR) is the ratio of the diagnoses in Long Island ZIP codes divided by the expected value calculated from New York State averages. Thus, it is still meaningful to compare our results to the state-level scan results. We identified clusters on Long Island that are exceptional compared to the state averages that went into the expected value calculation.

The evaluation of geographic relationships

The techniques employed by the local Moran and boundary techniques evaluate geographic relationships using the ZIP code and census geography of Long Island. In contrast, the scan technique, employed by New York State, uses ZIP code centroids and circles drawn around those centroids to represent Long Island's geography. Because of its long and forked appearance, the geography of Long Island is poorly represented by centroids and circles. How geographic relationships are evaluated is intimately linked to the description of geographic variation patterns and to cluster detection.

The Meaning of Geographic Variation

The term "clustering" by and of itself is so generic as to be almost meaningless for describing spatial variation in cancer morbidity. First, the differences between cluster detection methods already used to analyze Long Island data illustrate the notion of a cluster is meaningless without a precise description of the statistical test and its expectations. And second, an analysis approach that employs just one kind of cluster test is inappropriate. Obviously, a given technique will only detect the kinds of clusters it was designed to detect. In fact, experience has demonstrated that cancer morbidity evinces rich geographic variation, and we therefore should employ a variety of techniques to more fully describe relevant aspects of spatial pattern.

All Methods are Subjective

Because they are founded on assumptions and are more or less sensitive to different aspects of spatial pattern, all techniques for statistical pattern recognition are subjective, because they are founded on assumptions and are more or less sensitive to different aspects of spatial pattern. The spatial scan is based on a likelihood statistic, it uses centroids of areas to define spatial relationships, employs circular scan areas, and is univariate. Clearly, "clusters" can be multivariate, and not just univariate; they can be other shapes than circular; there are techniques other than centroids for evaluating spatial relationships; and likelihood is just one of several statistical approaches for identifying departures from a background morbidity. The point is not that the scan statistic is somehow flawed – in fact, it is one of the most powerful statistical techniques for identifying univariate clusters with the above-defined characteristics. Rather, the point is that there are many different aspects to spatial pattern. In order to explore these different aspects, researchers need to employ a variety of methods to more fully elucidate, characterize, and quantify the geography of cancer morbidity. The scan statistic is but one tool we can bring to bear on the study of geographic variation in cancer morbidity.

Conclusion

This is a methodological and comparative study to evaluate whether there is substantial benefit added by using a variety of techniques for geographic pattern detection at different spatial scales.

This paper demonstrates that there is significant spatial pattern of cancer in Nassau, Queens, and Suffolk counties on Long Island, New York. The general pattern that there are clusters of higher than expected cancer incidence on Long Island is consistent with other studies of the same area, though the exact locations and shapes of the clusters vary with the methods used by different researchers.

Several authors have observed that neither p-values nor confidence limits provide enough information to assess whether or not there is a true disease cluster caused by an environmental exposure. We advocate that the best approach would be to analyze the data using several cluster detection techniques, and if the area still sticks out like a sore thumb, there may be something there. As different tests identify different aspects of cluster morphology, the best characterization comes from an understanding of cluster shape, size, length, magnitude of excess, probability of occurrence, location of boundaries or gradients, relative locations of clusters and boundaries to each other, and finally correspondence of geographic patterns in health outcomes to potential exposures.

The differences in our results and that of the NYSDOH underscore the need for an exploratory, integrative, and multi-scalar approach to assessing geographic patterns of disease, as different methods identify different patterns. A concordance of the results from several different approaches increases the analyst's confidence that the suspected cluster indeed is unusual. By using several different methods – scan statistic, boundary analysis, local Moran – one is able to derive a more complete understanding of geographic variation in cancer morbidity on Long Island. One benefit is that researchers can now focus etiologic investigations at the finer spatial scales where local excesses in cancer morbidity are found on Long Island. Specifically, using a battery of approaches allows us to quantify different aspects of clusters; to explore different scales of clustering, and to evaluate how sensitive the results are to different definitions of clustering.

We recommend that future studies of geographic patterns use a concordance of evidence from a multiscalar integrative geographic approach to assure that 1) different aspects of spatial pattern are fully identified and 2) the results from the suite of analyses are logically consistent.

The obvious question after finding significant clusters of elevated and lower cancer incidence is – why? Could this

pattern be explained by environmental exposures, by differences in social factors, such as socio-economic status or ethnicity, or another factor that varies over this study region? We cannot exclude any particular explanation, as this study did not consider any economic, ethnic, or environmental exposure data. We do, however, consider one possible factor in the second paper in this set, airborne carcinogens [1]. But, as this is a study of encountered data, and data that are aggregated to a coarse spatial scale, it would be impossible to establish causation from these data on their own.

Authors' contributions

Authors GMJ and DAG collaborated intensely on all aspects of the manuscript, from research design to data preparation to presentation. Both authors wrote and approved the final manuscript.

Acknowledgements

We thank Dr. Ruth H. Allen, Environmental Epidemiologist and former US EPA Program Director for theLong Island Breast Cancer Study Project, Dr. Luc Anselin of the University of Illinois Urbana-Champaign, Dr. Dan Wartenberg, UMDNJ-RW Johnson Medical School, Piscataway, NJ, and Dr. Leah Estberg for suggestions, criticisms and comments that led to substantial improvements in the analysis and presentation. Dan Fagin of Newsday brought these data to our attention and encouraged us to undertake this analysis. The comments of Richard Hoskins, the co-editor of this journal, and three anonymous reviewers helped us improve the presentation of these results considerably. This study was partially funded by grant CA92669 from the National Cancer Institute (NCI). The opinions stated in this document are those of the authors and do not necessarily represent the official position of the NCI.

References

1. Jacquez GM and Greiling DA **Geographic boundaries in breast, lung and colorectal cancer in relation to exposure to our toxics in Long Island, New York.** *International Journal of Health Geographics* 2003, **2:**4
2. Kulldorff M, Feuer EJ, Miller BA and Freedman LS **Breast cancer clusters in Northeastern United States: a geographic analysis.** *Am J Epidemiol* 1997, **146:**161-70
3. Ries LAG, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg L and Edwards BK **SEER Cancer Statistics Review, 1973–1999.** *Bethesda, MD, National Cancer Institute* 2002, [http://seer.cancer.gov/csr/1973_1999/]
4. Kulldorff M and Nagarwalla N **Spatial disease clusters: detection and inference.** *Stat Med* 1995, **14:**799-810
5. Kulldorff M **A spatial scan statistic.** *Communications in Statistics – Theory and methods* 1997, **26:**1481-96
6. Anselin L **Local indicators of spatial association-LISA.** *Geographical Analysis* 1995, **27:**93-115
7. Rothman KJ and Greenland S **Chapter 14: Introduction to Categorical Statistics.** *In: Modern Epidemiology. Philadelphia: Lippincott-Raven Publishers* 1998,