International Journal of Health Geographics

Research

Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxics 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**:4

Received: 10 February 2003

Accepted: 17 February 2003

This article is available from: http://www.ij-healthgeographics.com/content/2/1/4

Abstract

Background: This two-part study employs several statistical techniques to evaluate the geographic distribution of breast cancer in females and colorectal and lung cancers in males and females in Nassau, Queens, and Suffolk counties, New York, USA. In this second paper, we compare patterns in standardized morbidity ratios (SMR values), calculated from New York State Department of Health (NYSDOH) data, to geographic patterns in overall predicted risk (OPR) from air toxics using exposures estimated in the USEPA National Air Toxics Assessment database.

Results: We identified significant geographic boundaries in SMR and OPR. We found little or no association between the SMR of colorectal and breast cancers and the OPR for each cancer from exposure to the air toxics. We did find boundaries in male and female lung cancer SMR and boundaries in lung cancer OPR to be closer to one another than expected.

Conclusion: While consistent with a causal relationship between air toxics and lung cancer incidence, the boundary analysis does not demonstrate the existence of a causal relationship. However, now that the areas of overlap between boundaries in lung cancer incidence and potential airborne exposures have been identified, we can begin to evaluate local- as well as large-scale determinants of lung cancer.

Background

This study is second in a two-paper series on cancer patterns on Long Island. In the first paper [1], we evaluated the spatial pattern of incidence of diagnoses of colorectal, breast, and lung cancers, identifying spatial clusters of high and low standardized morbidity ratio (SMR). In this paper, we compare cancer patterns to patterns in airborne carcinogens modeled in the National Air Toxics Assessment (NATA) database. While we acknowledge that environmental pollutant databases are imperfect and incomplete estimates of individual exposure, air toxics are one possible source of environmental exposure to carcinogens. If patterns in airborne toxins are significantly associated with cancer patterns, additional effort is warranted to determine whether or not there is a causative relationship. A more detailed understanding of spatial associations between patterns in health and environmental variables ultimately may lead to improved air quality and public health.

Health-environment relationships

Knowledge about possible relationships between human health and the environment is garnered in several ways. Laboratory studies explore how and whether toxic compounds cause disease at the organismic level. Epidemiological studies seek to identify whether *risk factors*, such as diet, socio-economic status and occupation, are associated with specific outcomes, such as breast cancer, in

human populations. With the advent of more detailed environmental data from remote sensing, toxic release inventories, and monitoring networks, the possibility of undertaking studies that relate geographic patterns in health outcomes to geographic patterns in environmental factors is now possible. Such studies seek to identify clusters of disease, and to relate the locations of those clusters to geographic patterns in environmental factors that might have caused the disease. Like other epidemiological studies, pattern analysis cannot establish causality. It can determine whether and where there is a statistical excess (or deficit) of disease, and whether locations of elevated disease are geographically associated with areas where plausible risk factors also are high. Susser and Susser [2] call for an integration of several levels of research, from the molecular, the individual, to the societal, because factors on each of these levels often interact to cause chronic disease.

Our purpose in undertaking this analysis is to illustrate how geographic pattern analysis can increase our understanding of breast, lung and colorectal cancers in Long Island, New York. Critical to this understanding is an appreciation of the assumptions, caveats, and limitations of the geographical approach and of this study in particular. While presented last, these considerations should be kept firmly in mind when making any kind of inference or decision from this study's results.

Geographic Pattern Analysis

Pattern recognition plays an important role in data summarization and description by identifying salient features and structure in the data. By asking a carefully crafted series of pattern analysis questions it is possible to evaluate specific hypotheses regarding the geographic patterns of disease in human populations. These hypotheses correspond to questions regarding *value*, *change* and *association*.

• *Value* questions have to do with the values of the variables surveyed, and how they are arranged in geographic space. Disease clustering is principally concerned with value questions such as "Is there an excess of disease?" and "Where are disease rates significantly high?"

• *Change* questions have to do with how values vary through geographic space and through time. Change questions include "Where do disease rates change rapidly?" and "Where do air toxics change rapidly in geographic space?"

• *Association* questions relate spatial pattern in one variable or set of variables to the pattern in another set of variables.

Example association questions include "Is spatial pattern in health outcomes associated with:

• The environment? (Environmental, occupational, and food-borne exposures),

• Population? (Demography, marriage, birth, ethnicity) and

• Individual? (Genetics, behavior, individual risk factors)?"

In this set of two studies we address three questions about breast, lung and colorectal cancers in Long Island.

1. Where are the statistically significant excesses and deficits of cancer? This value question is answered using disease clustering techniques in the first paper [1].

2. Where are the zones of rapid change (boundaries) in cancer incidence? This change question is answered using geographic boundary analysis.

3. Is geographic pattern in cancer incidence related to geographic patterns in carcinogen concentrations as modeled by the National Air Toxics Assessment program? This association question is evaluated using boundary overlap analysis.

Methods

Data

Cancer Incidence

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.

To protect patient privacy, the NYSDOH data provided case counts referenced to ZIP codes rather than individual residences. While ZIP codes are somewhat arbitrary spatial units of analysis with respect to potential health and environmental factors, they provide a convenient way to group the population and preserve confidentiality. We combined this dataset 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

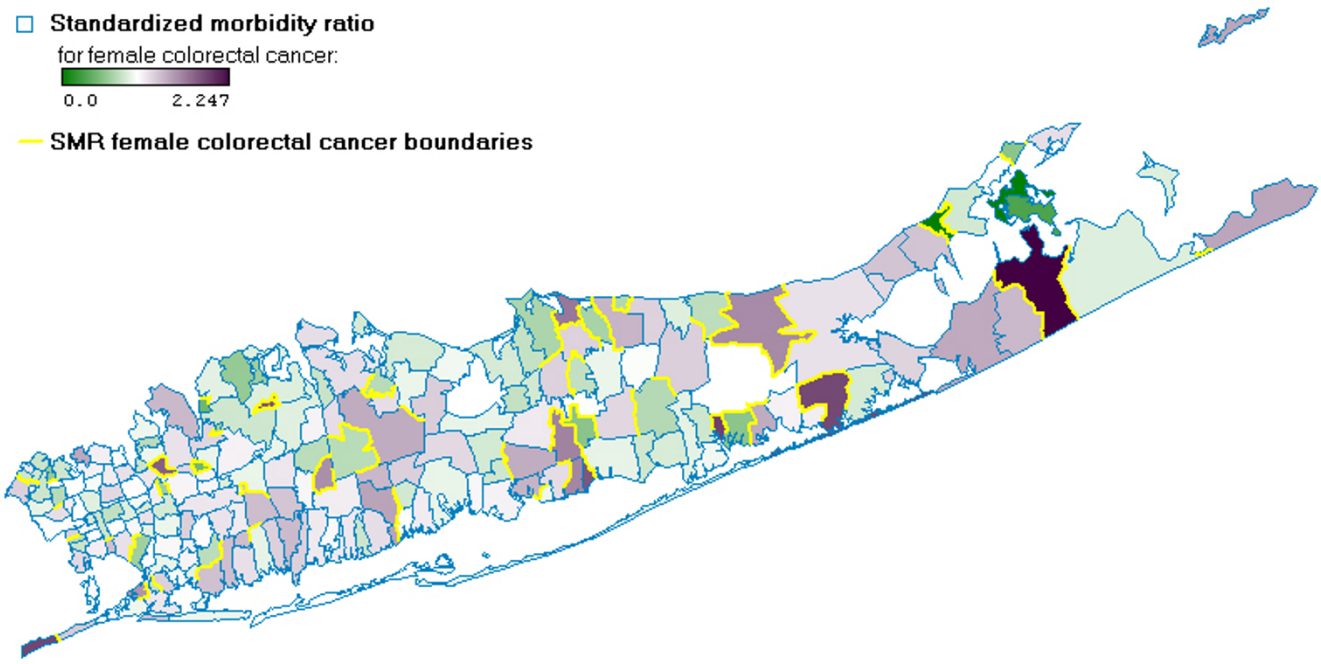

Figure 1
Colorectal cancer in females. The fill in the ZIP code areas indicates the SMR for female colorectal cancer, with darker purple regions having higher SMR, white regions having SMR near 1, and darker green regions having lower SMR. The boundaries shown in yellow indicate those ZIP code edges with large changes in cancer incidence. The blue outlines are the ZIP code edges.

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 long, 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.

National Air Toxics Assessment

The USEPA National Air Toxics Assessment (NATA, http://www.epa.gov/ttn/atw/nata/) combines information on point and nonpoint emissions of air toxics and weather information into an Assessment System for Population Exposure Nationwide (ASPEN). We obtained ASPEN model 1996 base-year data (Feb 2001 run). The exposure data is approximately concurrent with the cancer study period, thereby precluding any cause-and-effect interpretation, as cancers developed in 1993 could not have been caused by air toxics in 1996. Because of the latency in the development of cancer, it would not even be plausible to say that the 1996 data could explain only 1997 diagnoses. Yet, the 1996 data may be representative of the air toxics prior to 1996, and 1996 is the first year such a comprehensive geographic exposure model was available from the USEPA. As this is an opportunistic analysis, we took the data available. We thereby assume the 1996 data are reasonable representations of air pollution in the preceding decade during which causative exposures might have occurred. This assumption seems reasonable for air pollu-

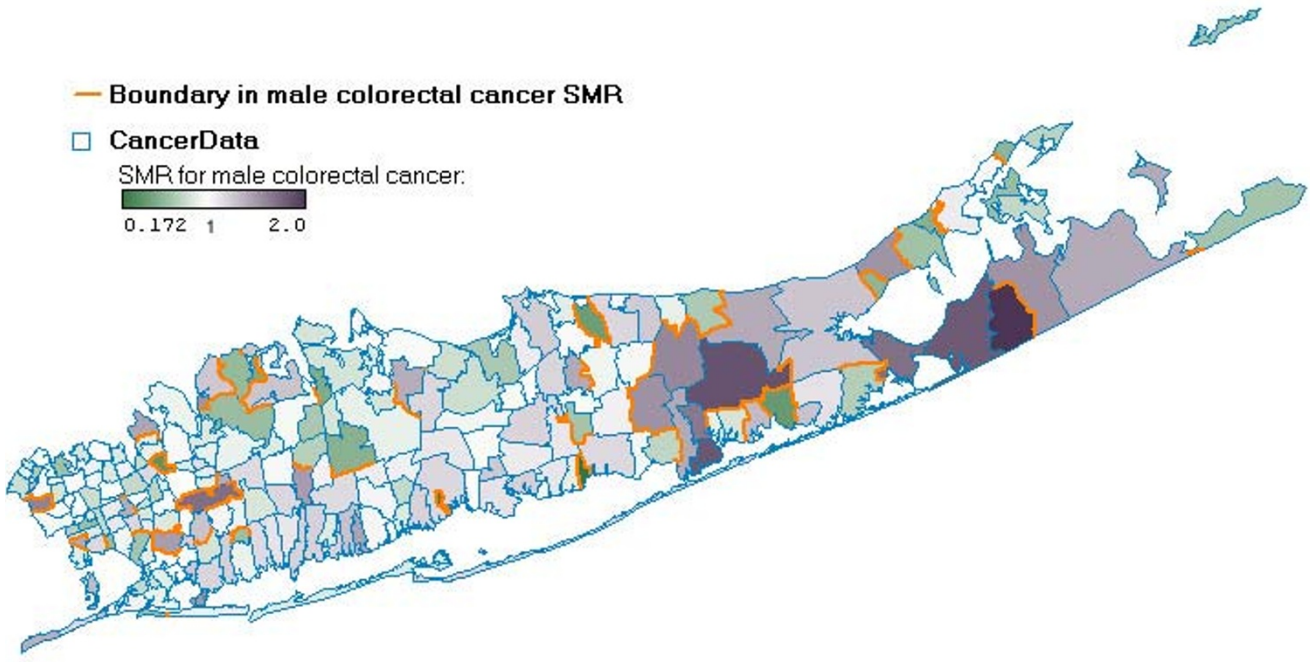

Figure 2
Colorectal cancer in males. The fill in the ZIP code areas indicates the SMR for male colorectal cancer, with darker purple regions having higher SMR, white regions having SMR near 1, and darker green regions having lower SMR. The boundaries shown in orange indicate those ZIP code edges with large changes in cancer incidence. The blue outlines are the ZIP code edges.

tion sources that have been in operation since the 1980s, and whose dispersal is mediated by transport mechanisms (e.g. prevailing winds) that haven't changed a great deal in the last 10–20 years. The ASPEN model estimates the average annual concentration of a series of known air toxics for all census tracts in the nation. We used concentrations of only those air toxics thought to be potential carcinogens for the three study cancers (Table 7). This list by no means constitutes an exhaustive list of potential carcinogens on Long Island. For the purposes of this study, the compounds in this list were deemed the most plausible carcinogens and exposure to these compounds was combined into a single risk measure, the overall predicted risk (OPR), defined below (Equation 2).

As exposure to each compound has a different risk, we standardized the exposure by multiplying the estimated average annual concentration of each compound by its Unit Risk Estimate (URE) as shown in Equation 1. The URE is the lifetime risk of excess cancer cases predicted to come from continuous exposure to a compound at a concentration of 1 $\mu g/m^3$ in the air (for more information see definition on the NATA website, http://www.epa.gov/ttn/atw/nata/gloss1.html. UREs may under- or over-estimate the actual risk of exposure to these compounds, as the predictions are extrapolations from tests in animals and/or the effects of low doses. All UREs are from the Draft USEPA NATA report [5], except that for diesel particulate matter. The USEPA has not yet defined a URE for diesel and so we used the midpoint of the URE range from the California EPA [6]. To calculate exposure for each compound we used the following formula:

$$Exposure \times URE = CancerRisk \quad \text{(Equation 1)}$$

We obtained the annual estimated exposure from the NATA dataset, and used the URE values from Table 7 to obtain estimates of excess cancer cases due to that exposure. As the URE is a risk estimate for all cancer, rather than a cancer-specific figure, the OPR for each cancer is likely an overestimate of risk for an individual cancer.

The use of a national-scale assessment to predict cancer risk based on air toxics is subject to caveats that have been identified by the EPA:

Table 1: Colorectal cancer, subboundary statistics.

Boundaries	Statistic	Observed	Expected	P↑	P↓
Female colorectal cancer	Ns	46	37	0.032	0.998
	Lmean	2.130	2.687	0.988	0.032
	Lmax	7.000	12.257	0.996	0.048
Male colorectal cancer	Ns	39	43	0.860	0.184
	Lmean	2.513	2.286	0.184	0.860
	Lmax	10.000	10.313	0.528	0.628
OPR for colorectal cancer	Ns	123	266	1.000	0.004
	Lmean	5.407	2.504	0.004	1.000
	Lmax	60.000	17.406	0.004	1.000

In this and all subsequent tables, P↑ indicates the upper tail p-value and P↓ indicates the lower tail p-value.

Table 2: Colorectal cancer, overlap statistics.

Boundaries overlapped	Statistic	Observed (meters)	Expected (meters)	P↑	P↓
Male (g) and female (h) colorectal cancer boundaries	Og	1166.768	1971.987	0.980	0.024
	Oh	1258.268	1718.560	0.852	0.152
	Ogh	1212.518	1845.364	0.988	0.016
	Os	58	43.916	0.012	0.992
Female colorectal cancer boundaries (g) and OPR boundaries (h)	Og	13508.245	775.437	0.004	1.000
	Oh	2052.526	2029.171	0.380	0.624
	Ogh	3531.657	1867.316	0.004	1.000
	Os	0.000	0.104	1.000	0.900
Male colorectal cancer boundaries (g) and OPR boundaries (h)	Og	12751.844	770.816	0.004	1.000
	Oh	2034.478	1834.967	0.152	0.852
	Ogh	3418.275	1697.481	0.004	1.000
	Os	0.000	0.129	1.000	0.888

"The UREs used in the national-scale assessment are subject to four major areas of variability and uncertainty. First, many of the pollutants were classified as probable carcinogens because data were not sufficient to prove causality in humans. It is possible that some of these pollutants do not cause cancer at environmentally relevant doses, and that true risk associated with these air toxics is zero. Second, all UREs in this study were based on linear extrapolation from high to low doses. It is possible that the true dose response relationships for some pollutants may be less than linear, resulting in an overestimate of risk. Third, most UREs in this study were developed from animal data using conservative methods to extrapolate between species. Human responses may differ from the predicted ones. The first three elements are comprised entirely of uncertainty. Fourth, most UREs in this study were based on statistical upper confidence limits, though some were based on statistical best fits. (While this does not affect overall uncertainty, UREs based on best fits should be unbiased, while those based on upper confi-

Table 3: Breast cancer, subboundary statistics.

Boundaries	Statistic	Observed	Expected	P↑	P↓
Female breast cancer	Ns	46	42.518	0.276	0.800
	Lmean	2.130	2.334	0.800	0.276
	Lmax	8.000	10.305	0.828	0.308
OPR for breast cancer	Ns	99	270.446	1.000	0.004
	Lmean	6.707	2.460	0.004	1.000
	Lmax	192	16.574	0.004	1.000

dence limits should be biased high.) This fourth element represents a combination of variability (i.e., based on variation responses of different people or animals) and uncertainty (i.e., potential errors in the measurement of exposure and response). Because of the aggregate treatment all four sources of variability and uncertainty described above, EPA considers all its UREs to be upper-bound estimates."

-pages 112–3, http://www.epa.gov/ttn/atw/sab/natarepport.pdf

Regarding the use of UREs, one should note that the methods employed are sensitive to the relative, rather than absolute value, of the risk estimates. By focusing on boundaries, we are able to identify spatial structure and geographic associations so long as the relative values of the OPRs are correct. Hence the methods employed will yield the same results for a biased risk estimator – provided the bias on average is the same for all observed values. We also note the UREs in Table 7 are for all cancers combined, and not for site-specific cancers. Our analysis identified compounds thought to be carcinogens for each of the 5 site-specific cancers we considered, and then calculated a site-specific URE based on the values in Table 7. Ideally, one should use site-specific UREs, but these are not yet available from EPA.

We calculated an overall predicted risk from air toxics (OPR) for each cancer by summing up the excess cancer cases for each of the relevant compounds as shown in Equation 2.

$$\Sigma\ CancerRisk = OPR \quad \text{(Equation 2)}$$

Summing up the excess cancer cases for all of the relevant compounds assumes an additive relationship – that particular compounds do not interact in a synergistic or threshold-related manner to influence dose-response relationships. The EPA is currently using an additive model for assessing dose and response to multiple compounds, but further research needs to be done to confirm the additive model or else replace it with a more appropriate model. Again, pattern recognition approaches are useful under this kind of uncertainty since they are relatively robust provided the rank order of the estimates is "about right."

Local Boundary and Subboundary Analysis

Borders where SMRs change a great deal may indicate areas where causative exposures change through geographic space, where SMRs are unstable, and/or where local populations differing in cancer incidence abut. The identification of such borders may provide insight into the causes, correlates and uncertainties in cancer incidence. To detect local boundaries we used the Womble [5] approach. Wombling identifies those locations with the highest local rates of change (measured by squared Euclidean distance between SMR values in adjacent ZIP codes). We used a gradient value threshold of 20%, so the top 20% of all local rates of change in the dataset were called boundaries. Wombling has been applied to raster data [6–8] and point data [9,10]. It was extended to polygon data by Susan Maruca and Geoffrey Jacquez in the BoundarySeer software http://www.terraseer.com/boundaryseer.html. To our knowledge, this publication is the first application of this new wombling approach.

Because choosing a boundary threshold value is subjective, we evaluated the boundaries detected statistically through subboundary analysis [11]. For each defined set of boundaries we calculated the number of singletons Ns, mean boundary length Lmean, mean maximum boundary length Lmax, mean boundary diameter, and mean maximum boundary diameter. We will report Ns, Lmean, and Lmax. We then evaluated the probability of the observed value of each subboundary statistic against the null hypothesis of no spatial structure in the underlying variable (either SMR or OPR) through Monte Carlo randomizations. In these randomizations, the observed SMR values were randomized across the ZIP codes of Long Island. With equations 3 & 4, these statistics can be evaluated as excessively high (significant upper tail p-value or P↑,

when the observed value is significantly higher than those in the reference distribution) or as excessively low (significant lower tail p-value or P↓, when the observed value is significantly lower than those in the reference distribution). Thus the statistics can be interpreted to identify statistical evidence of boundary cohesiveness (longer boundaries than expected by chance high Lmean and Lmax and low Ns) or fragmentation (shorter boundaries than expected by chance low Lmean and Lmax and low Ns – P↑).

Boundary Overlap Analysis

To assess the association between two sets of boundaries (e.g. cancer incidence (SMR) boundaries and cancer risk (OPR) boundaries) we used boundary overlap statistics [12]. We evaluated four statistics of boundary overlap based on the average minimum distance from boundaries in one variable (e.g. SMR) to the nearest boundary in the other variable (e.g. OPR). They are Os, Og, Ogh, and Oh. Os is the count of the number of boundary locations that are included in both sets of boundaries. Og is the mean distance from the boundaries of one variable (g) to the nearest boundary location in another boundary set (h). Oh is the mean distance from h to the nearest point in g. Ogh is the mean distance from locations in either boundary to the nearest location in the other

We obtained a p-value through equations 3 & 4 for the observed overlap by comparing the observed values of all four statistics to those generated by Monte Carlo randomizations. BoundarySeer randomized the variables considered (SMRs and/or OPR), recalculated the boundaries, and then recalculated the overlap statistics. The null hypothesis for this randomization approach is that boundaries in cancer are independent (not associated) with boundaries in cancer risk. Like the subboundary statistics, these overlap statistics can be evaluated as significantly closer (high Os, low Og, Oh, or Ogh) or significantly farther than expected by chance (low Os, high Og, Oh, or Ogh).

Because the SMR and OPR data were assigned to different geographic units (ZIP codes and census tracts, respectively) we would almost never expect overlap of SMR boundaries and OPR boundaries to result in significantly high Os, as Os depends on the exact location of the boundaries. The other statistics are minimum distances, and so are more reasonable measures of coincidence of two sets of boundaries detected on different geographic units (e.g. census tracts vs. ZIP codes).

Calculating p-values

Upper and lower p-values provide a sense of how extreme the observed values of the subboundary and overlap statistics are compared to the reference distribution of values obtained by randomization. The formulae for calculating these p-values are:

$$P\uparrow = \frac{NGE + 1}{Nruns + 1} \quad \text{(equation 3)}$$

$$P\downarrow = \frac{NLE + 1}{Nruns + 1} \quad \text{(equation 4)}$$

where Nruns is the total number of Monte Carlo simulations, NGE is the number of simulations greater than or equal to the observed value of the statistic, and NLE is the number of simulations less than or equal to the observed value of the statistic.

Results

Colorectal Cancer

Females

Boundaries in female colorectal cancer are shown in Figure 1. Our analysis identified those ZIP code edges where cancer incidence changes the most. In general, the boundaries in female colorectal cancer circumscribe or partially surround only one ZIP code. This pattern is consistent with the smaller-scale clustering found for females relative to males under the local Moran test [1]. For example, there were 3 clusters of female colorectal cancer, each comprised of from 3 to 4 ZIP codes [1, Table 1], vs. 5 clusters of male colorectal cancer, each comprised of from 3 to 7 ZIP codes [1, Table 2]. These smaller clusters indicate that geographic clustering of female colorectal cancer occurs at a smaller spatial scale than for males. This finding is further substantiated by subboundary analysis (Table 1), which found the boundaries in female colorectal cancer to be significantly fragmented. There are significantly more singleton boundaries (Ns larger than expected, P↑ = 0.032), while the boundaries are shorter (significantly low Lmean, P↓ = 0.032) than is expected under this null hypothesis.

Males

Boundaries in male colorectal cancer are shown in Figure 2, and identify margins of ZIP codes that differ substantially in cancer incidence from their neighbors. These indicate not only the zones of high variation in cancer incidence that are expected at the margins of the significant clusters occurring under the local Moran analysis [1], but also highly local boundaries indicative of spatial variation in incidence at small spatial scales. Several boundaries appear to be long, and connect several ZIP codes; others are quite short and are comprised only of part of the margin of a ZIP code. Under subboundary analysis (Table 1), the boundaries, as a whole, were found to be neither significantly long nor significantly fragmented. We believe this result is consistent with an overall pattern of boundary fragmentation in western Long Island, and

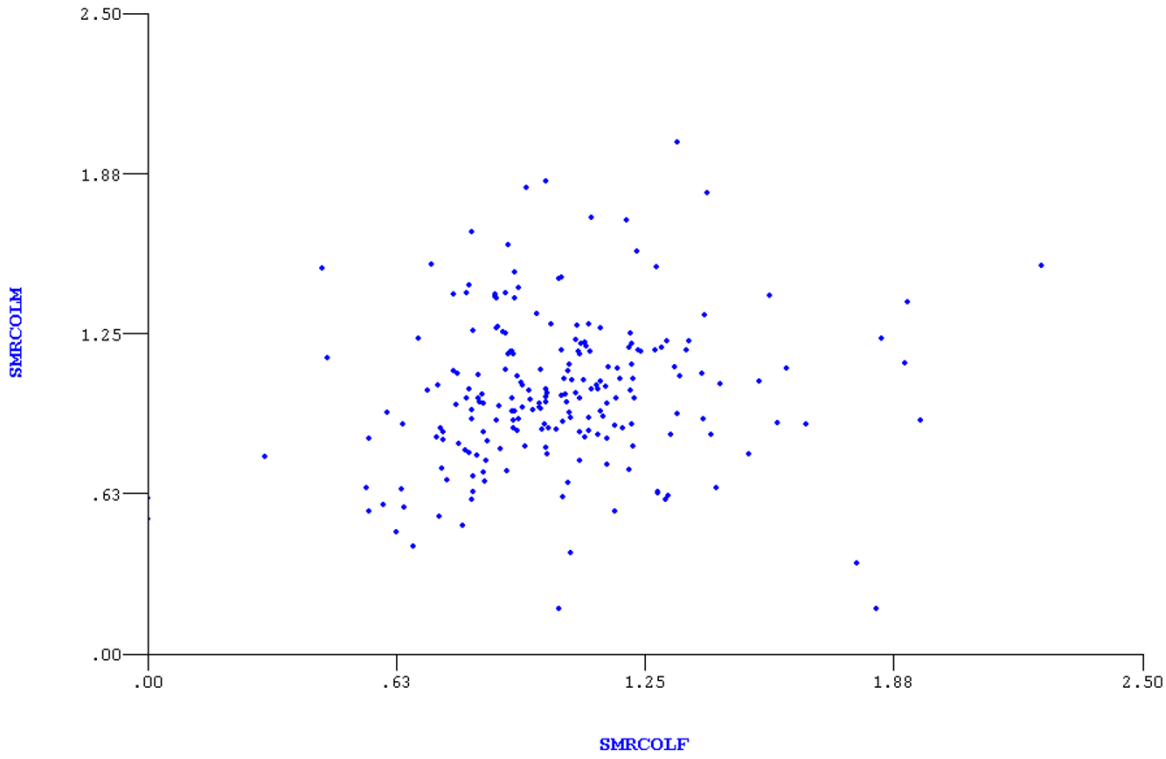

Figure 3
Scatter plot of the incidence of male (y-axis) versus female (x-axis) colorectal cancer. While there is a significant relationship between male and female colorectal cancer ($p < 0.001$) that relationship explains only 2.5% of the variation in the data, hardly a compelling explanation.

cohesive boundaries around the large-scale clusters occurring on mid- to eastern Long Island. This finding suggests that the determinants of colorectal cancer operate on increasingly larger spatial scales as one moves from west to east.

Males and females

Colorectal cancer has both dietary and genetic determinants, in addition to other risk factors such as age and smoking. Because diet is strongly influenced by the family environment, one might expect the incidence of male and female colorectal cancer to covary. To explore this expectation we generated bivariate plots of male vs. female cancer incidence (Figure 3), and also conducted a boundary overlap analysis. The scatter plot suggests little, if any, association between male and female colorectal cancer incidence.

Now consider the map of male and female colorectal boundaries (Figure 4). In several areas the female colorectal boundaries (yellow) overlap the male colorectal boundaries (orange) exactly, displayed as a yellow line with orange margins. These boundaries have significant exact overlap (Os $P\uparrow = 0.012$, Table 2). Further, the average minimum distance between the male and female colorectal boundaries was significantly smaller than expected (Ogh $P\downarrow = 0.016$). While the locations of male boundaries in colorectal cancer tended, on average, to occur near boundaries in female colorectal cancer (Og $P\downarrow = 0.024$) boundaries in female colorectal cancer were not

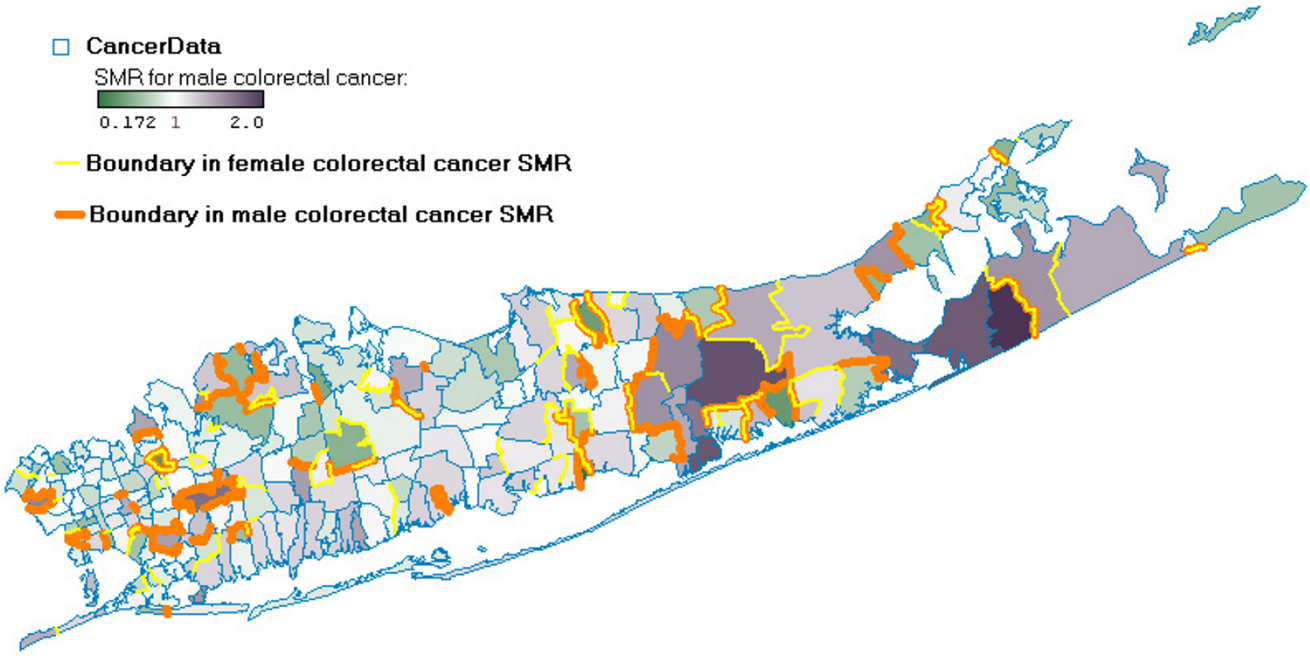

Figure 4
Map of male and female colorectal incidence and boundaries. This map shows the boundaries for male and female colorectal cancer superimposed on the map of male colorectal SMR. The blue lines are ZIP code edges. The boundaries shown as thin yellow lines indicate those ZIP codes with large changes in female colorectal cancer incidence. The boundaries shown in thick orange lines indicate the edges of ZIP codes across which there are large changes in male colorectal cancer incidence. Locations where the thick orange line surrounds the yellow line illustrate exact boundary overlap.

necessarily near boundaries in male colorectal cancer (Oh $P\downarrow = 0.152$). These results are indicative of the ability of overlap statistics to detect common spatial variation patterns even when the observed correlation between two variables is weak. Despite the substantial 'noise' in the plot of male vs. female colorectal cancer incidence (Figure 3), overlap analysis revealed the geographic association expected given common exposures among males and females or attributable to similar diet and genetics within family units.

Colorectal Cancer – Analysis of NATA Data

The overall predicted risk for colorectal cancers was calculated from the NATA dataset and mapped (Figure 5). There is an outlier of high risk in census tract 11200, in the vicinity of Jamaica, Ozone Park, ZIP code 11416. Whether this is attributable to small population size, reporting differences, or other causes was not further explored in this study.

Subboundary analysis

Figure 6 is the map of boundaries in OPR for colorectal cancer, overlaid with boundaries in colorectal cancer risk (OPR) and boundaries in the incidence of male and female colorectal cancer. The boundaries in OPR are significant and cohesive, being longer and having fewer singletons than expected under the null hypothesis. The number of subboundaries were significantly fewer than expected (Ns $P\downarrow = 0.004$, Table 1). The boundaries were significantly long (Lmean $P\uparrow = 0.004$; Lmax $P\uparrow = 0.004$). In total, these results indicate boundaries that are significantly longer and more cohesive than is expected by chance, and suggests that spatial variation in OPR for colorectal cancer occurs on relatively large spatial scales. This outcome is consistent with the model used by EPA that incorporates both point- and area-sources into the air quality model.

Overlap analysis

Boundary overlap analysis determined whether zones of rapid change in OPR are significantly associated with boundaries in colorectal cancer SMR. If the air toxics modeled in the NATA database are indeed strongly associated with colorectal cancer risk, then we would expect boundaries in OPR to be significantly close to boundaries in colorectal cancer SMR. Accordingly, an overlap analysis was undertaken for both male and female colorectal cancers. For female colorectal cancer SMR we found overlap

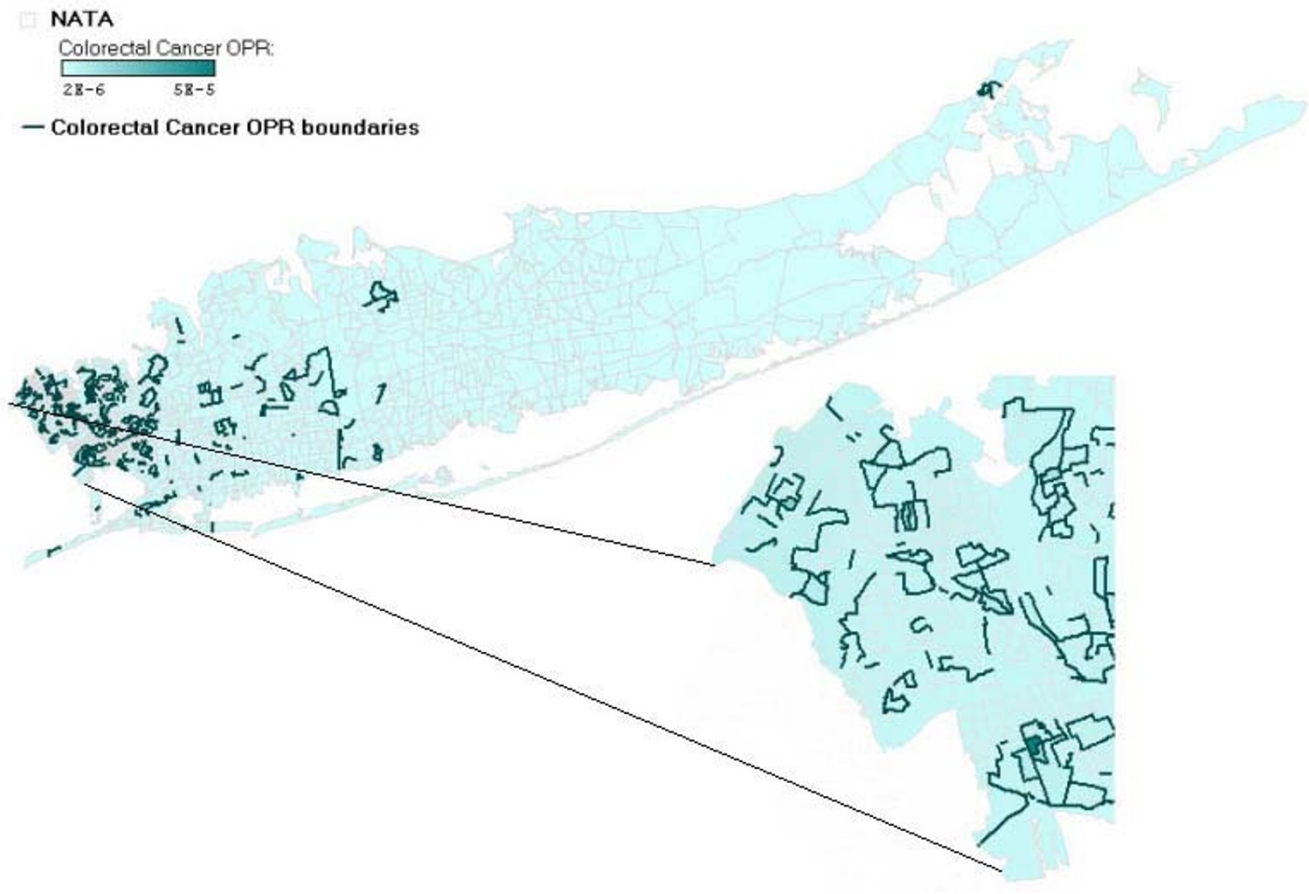

Figure 5
Geographic distribution of overall predicted risk for colorectal cancer. The turquoise fill indicates the OPR for colorectal cancer, with darker regions having higher OPR, the gray lines outline the census tract edges. The dark green lines indicate boundaries in colorectal cancer OPR. The color gradient in this map is influenced strongly by the outlier in the western part of long island (shown in the inset).

avoidance: boundaries in female colorectal cancer incidence are further from boundaries in colorectal OPR than is expected by chance (Table 2). Taking Long Island as a whole, the average minimum distance from a boundary in female colorectal cancer SMR to the nearest boundary in colorectal OPR is significantly larger (Og $P\uparrow = 0.004$) than its expected value. The same result obtains for males, where the average minimum distance from a boundary in male colorectal cancer SMR to the nearest boundary in colorectal is significantly larger (Og $P\uparrow = 0.004$) than its expected value.

When considering these statistical results with the maps in Figures 1 and 2, the source of overlap avoidance is apparent. The majority of boundaries in OPR are found in Western Long Island, in urban areas, suggesting that a greater number of point source emissions are causing greater spatial variation in OPR in more urban areas. In fact, the eastern-most boundary in colorectal OPR occurs in the vicinity of Greenlawn (ZIP 11740), while boundaries in both male and female colorectal cancer SMR are found as far east as Wainscott and Fishers Island. Thus while overlap avoidance occurs on the scale of Long Island as a whole, further investigation is needed to evaluate whether overlap occurs specifically in urban areas where there is a great deal of local scale variation in colorectal OPR.

Breast Cancer

We conducted a local boundary analysis (Figure 7) to identify the edges of ZIP codes where female breast cancer incidence changes rapidly. As a group, the breast cancer SMR subboundaries are not statistically remarkable. The number of subboundaries is near its expectation (Ns $P\downarrow$ =

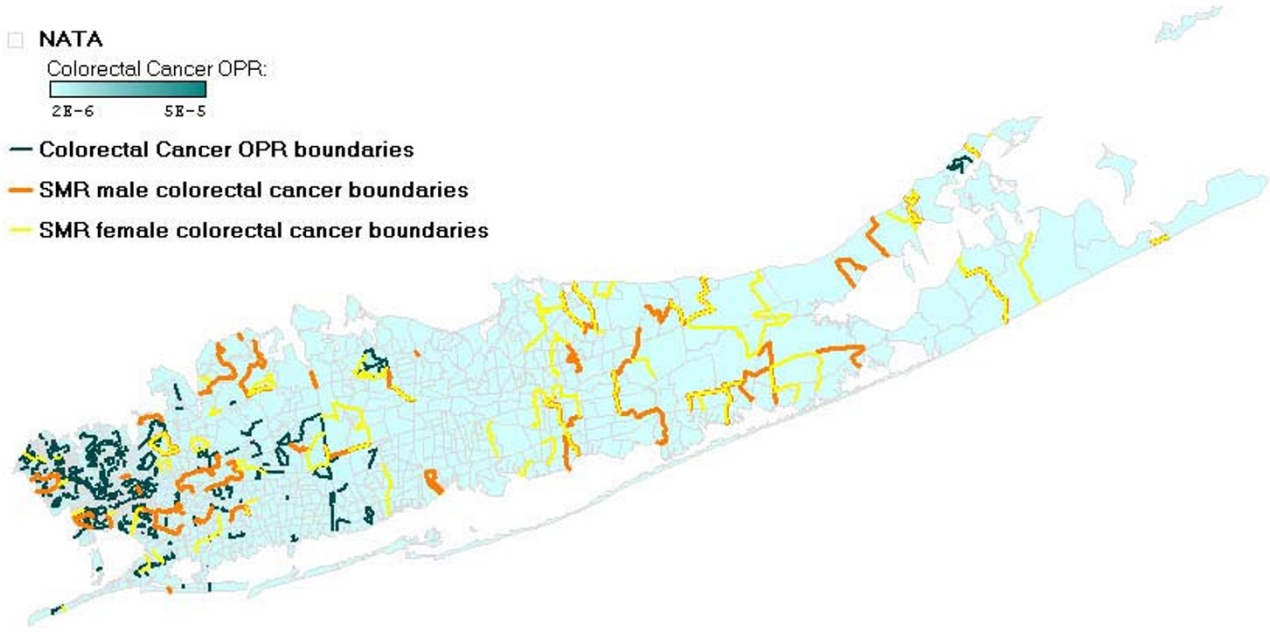

Figure 6
Map of overall predicted risk (OPR) for colorectal cancer, overlaid with boundaries in OPR and SMR. The fill color in the census tracts represents the OPR for colorectal cancer. This map also shows boundaries in OPR for colorectal cancer (dark green) and in male (orange) and female (yellow) colorectal cancer SMR. The gray lines define the census tract edges. The turquoise fill indicates the OPR for colorectal cancer, with darker regions having higher OPR. Boundaries in male and female colorectal cancer overlap significantly. Boundaries in OPR and both male and female colorectal cancers repel to a significant extent, primarily because geographic variation in OPR is concentrated in Western urban areas.

0.800). These subboundaries have lengths near what would be expected by chance (Lmean $P\uparrow = 0.800$,).

We found spatial pattern in breast cancer SMR at two distinct spatial scales. The first scale is at the level of the individual ZIP code, resulting in the spatial outliers described earlier [1]. The second scale occurs across adjacent ZIP codes, resulting in clusters of three or more ZIP codes found under the local Moran test [1]. The spatial scale of the pattern gives us some insights into the likely spatial scale of the generating process. For example, it seems unlikely that spatial outliers in SMR would result from a carcinogen, such as an airborne toxic, dispersed over a large geographic area. At the same time, we wouldn't expect a cluster of several counties to result from a highly localized exposure. Thus, if we are to use spatial pattern in breast cancer SMR as a clue to underlying causative exposures, we will need to consider exposure mechanisms that operate at both small (sub ZIP code) and local (bridging 3–7 ZIP codes) spatial scales.

Breast Cancer – Analysis of NATA Data

The overall predicted risk (OPR) for breast cancer was calculated from the NATA data set and mapped (Figure 8). We see a broad area of moderate to low overall predicted risk extending from west central to far eastern Long Island. Areas of moderate to high overall predicted risk are found in the western urban areas, especially in the vicinity of East Elmhurst, Maspeth, Long Island City and Little Neck in Flushing.

Subboundary Analysis

The boundaries in breast cancer OPR are significant and cohesive, being longer and having fewer singleton boundaries than expected by chance (Table 1). The number of subboundaries is significantly fewer than expected (Ns $P\downarrow = 0.004$). Both the mean and maximum boundary length were longer, on average, than expected by chance (Lmean $P\uparrow = 0.004$; Lmax $P\uparrow = 0.004$). These results indicate boundaries that are significantly longer and more cohesive than is expected by chance.

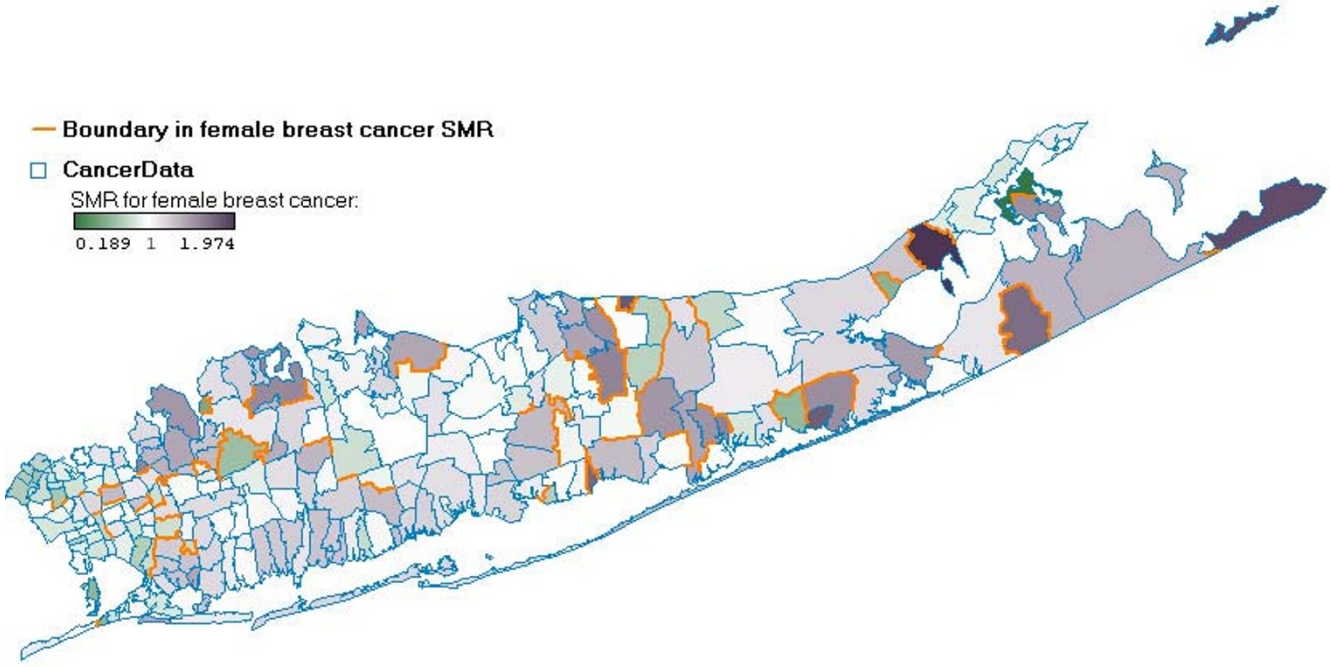

Figure 7
Female breast cancer boundaries. The blue outlines are the ZIP code edges. The fill in the ZIP code areas indicates the SMR for female breast cancer, with darker purple regions having higher SMR, white regions having SMR near 1, and darker green regions having lower SMR. The boundaries shown in orange indicate those ZIP code edges with large changes in cancer incidence.

Overlap Analysis

Comparing Figure 9 to Figure 3 in [1], we located several areas of high OPR for breast cancer near the cluster of low breast cancer incidence identified in the local Moran analysis [1]. We also note the cluster of high breast cancer SMR found on southeastern Long Island is in an area of low OPR [1]. Based on map inspection, overall there appears to be a negative relationship between OPR and breast cancer incidence so that clusters of high breast cancer incidence occur where OPR is low, and clusters of low incidence occur where OPR is high. This result based on visual inspection is supported by the statistical analysis of boundary overlap (below).

Boundary overlap analysis was used to determine whether boundaries in breast cancer OPR were closer to boundaries in breast cancer SMR than one would expect were these variables independent. If breast cancer SMR indeed is increased in those locations where the airborne toxics underlying the OPR for breast cancer are elevated, then we should expect OPR and breast cancer SMR to have similar spatial patterns, and their boundaries should be significantly close to one another. We undertook a boundary overlap analysis to evaluate this hypothesis. Taking Long Island as a whole, the boundaries in breast cancer SMR are further away from boundaries in OPR than one would expect by chance (Figure 9, Table 4). The observed average minimum distance from a boundary in breast cancer SMR to a boundary in OPR was significantly higher than expected (Og $P^{\uparrow} = 0.004$). Thus boundaries in breast cancer avoided boundaries in OPR such that one tends to find boundaries in breast cancer in locations where there aren't boundaries in OPR. This finding is consistent with the observed locations for the larger, multiple ZIP code breast cancer SMR clusters found under with the local Moran test [1], and the singleton clusters found by the local boundary analysis. There thus appears to be boundary avoidance, so that zones of rapid change in breast cancer incidence aren't found near zones of rapid change in OPR; and an inverse relationship between OPR and SMR, so that high values of breast cancer incidence tend not to be found where OPR is high.

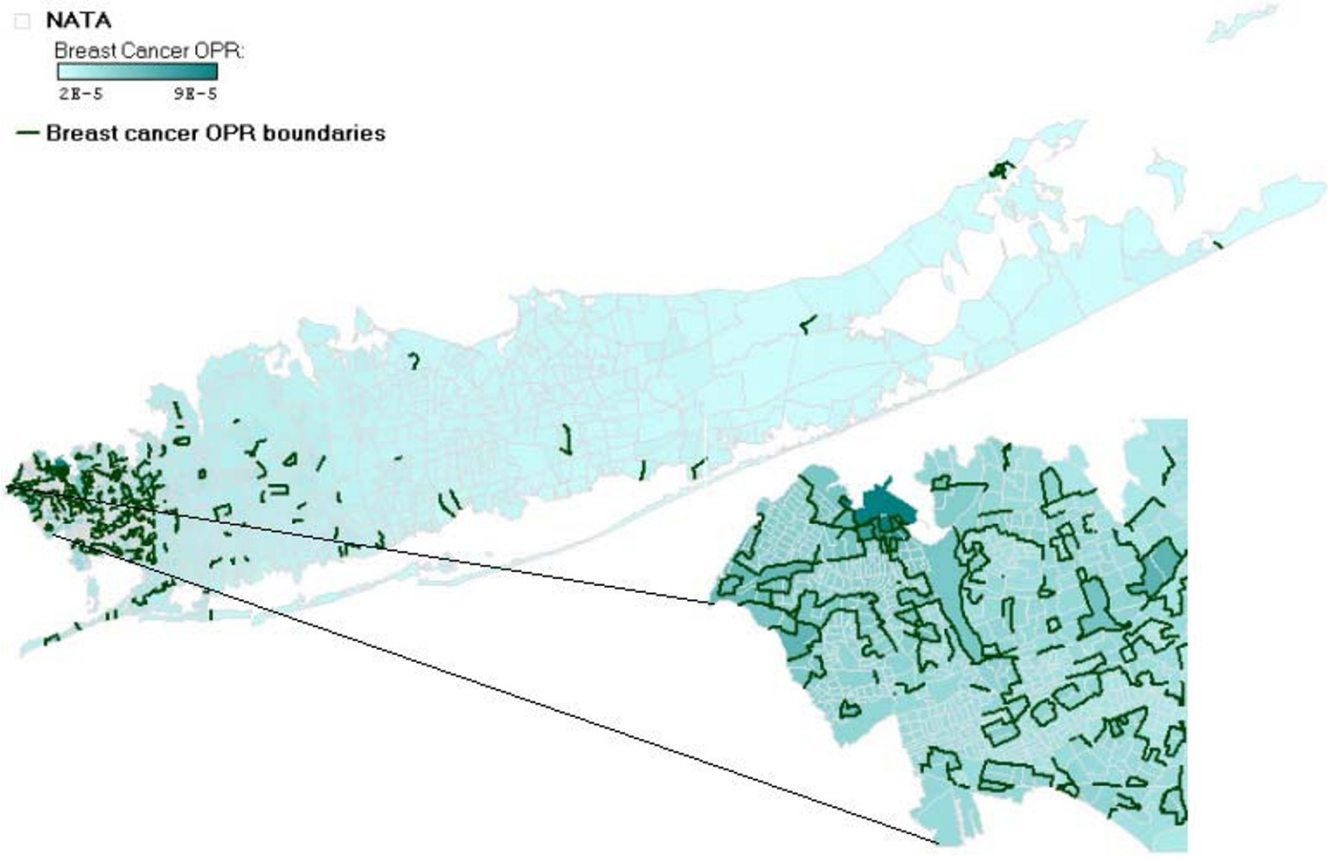

Figure 8
Geographic distribution of overall predicted risk (OPR) for breast cancer. The turquoise fill indicates the OPR for each census tract (outlined in gray). Darker regions have higher OPR than lighter regions. The boundaries in breast cancer OPR are shown in dark green.

Lung Cancer

Females

Boundaries in female lung cancer SMR are shown in Figure 10. These boundaries are the margins of abutting ZIP codes with very different lung cancer SMR values. When we consider Long Island as a whole, these boundaries are neither significantly fragmented nor contiguous. Indeed each of the subboundary statistics we examined – the number of singleton boundaries (Ns) and boundary mean and minimum length (Lmean, Lmax) – were not significantly different from the values expected under the null hypothesis of a random geographic distribution of SMR values (Table 5). This implies that when Long Island is considered as a whole, spatial processes that lead to boundary generation or to boundary fragmentation are either absent or countervailing.

Males

Boundaries in male lung cancer incidence are shown in Figure 11. These are expected to indicate not only the margins of the clusters identified under the local Moran test [1], but also the margins of singleton ZIP codes with SMR values that differ substantially from their neighbors. It is interesting to observe the high singleton ZIP codes (Rockaways, Brightwaters and Roosevelt) are along the southern shore.

Like female lung cancer incidence, the boundaries in male lung cancer appear to be a mixture of long boundaries demarcating the large-scale clusters identified under the local Moran test, and smaller-scale boundaries circumscribing singleton ZIP codes whose SMRs differ substantially from their immediate neighbors. Not surprisingly, when Long Island is considered in its entirety the boundaries in lung male cancer incidence are neither significantly long nor significantly fragmented under sub-

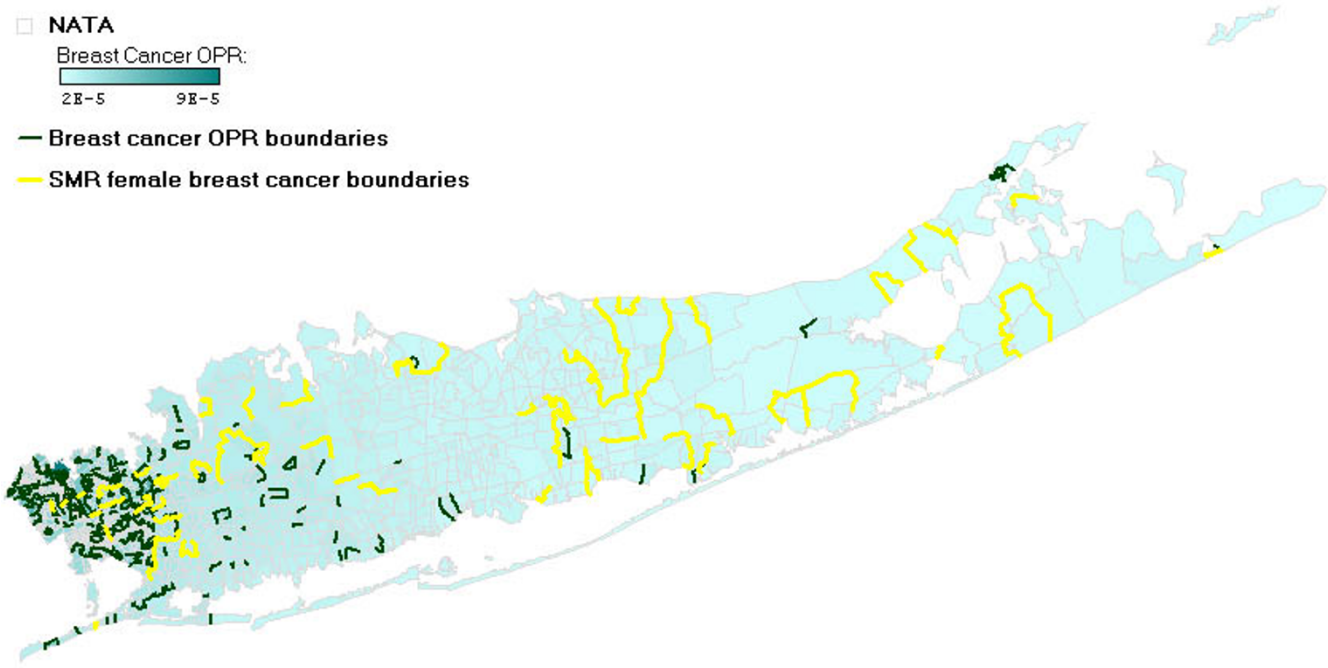

Figure 9
Map of overall predicted risk (OPR) and boundaries for breast cancer. The fill in the census tracts represent OPR, overlaid with boundaries in OPR for breast cancer (dark green), and boundaries in breast cancer incidence (yellow). The gray outlines are the census tract edges. Boundaries in OPR and breast cancer repel to a significant extent.

Table 4: Breast cancer, overlap statistics.

Boundaries overlapped	Statistic	Observed (meters)	Expected (meters)	P↑	P↓
Female breast cancer boundaries (g) and OPR boundaries (h)	Og	4313.784	725.902	0.004	1.000
	Oh	1909.643	1864.621	0.364	0.640
	Ogh	2219.244	1718.293	0.036	0.968
	Os	0.000	0.116	1.000	0.888

boundary analysis (Table 5). Because both spatial outliers ZIP codes as well as large-scale clusters were found, this result suggests that boundary-generating processes are acting at both local (within ZIP codes) and large (across several ZIP codes) geographic scales.

Females and males

It is known that smoking is a major determinant of lung cancer and that smoking as a behavior tends to cluster in families and is associated with certain socio-demographic groups. In addition, even when only one household member smokes, the risk for lung cancer is elevated among other household members through "second hand" smoking. We thus might expect to find a correlation between male and female lung cancer incidence. This was explored via a scatter gram and by boundary overlap analysis. The scatter gram (Figure 12) suggests a weak but positive correlation between male and female lung cancer SMR.

Boundary overlap (Figure 13) revealed that there is some overlap in male and female lung cancer boundaries, espe-

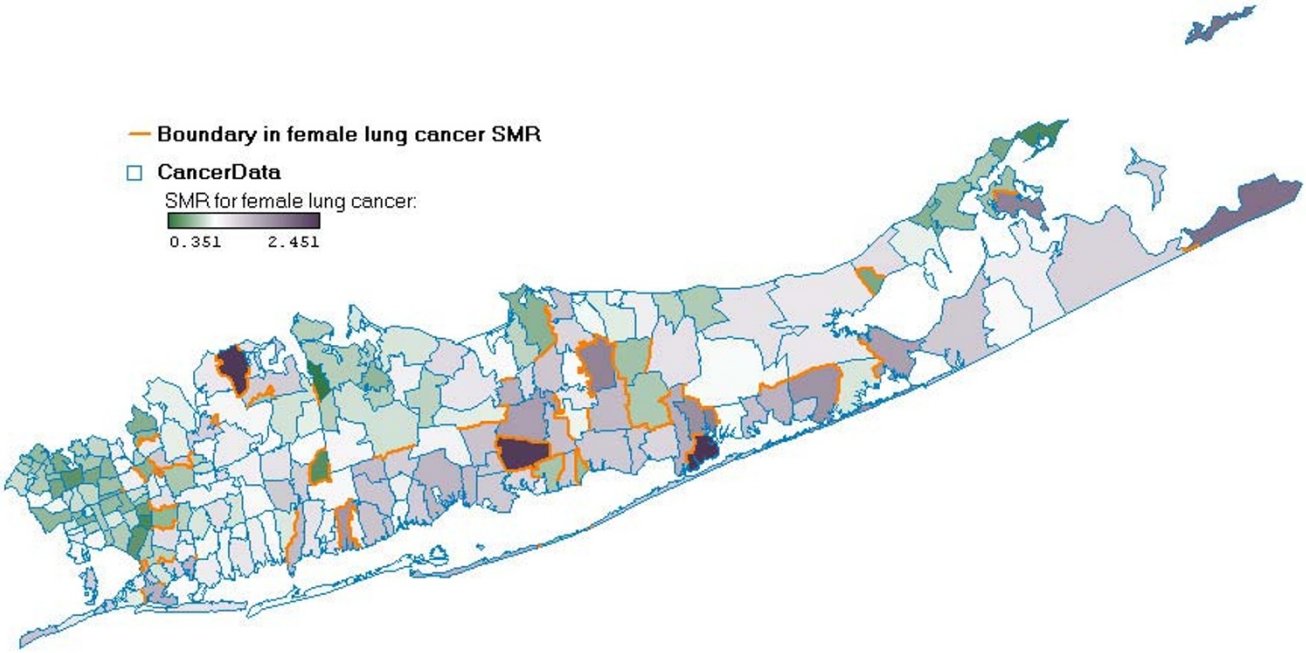

Figure 10
Female lung cancer showing local boundaries. The blue outlines are the ZIP code edges. The fill in the ZIP code areas indicates the SMR for female lung cancer, with darker purple regions having higher SMR, white regions having SMR near 1, and darker green regions having lower SMR. The boundaries shown in orange indicate those ZIP code edges with large changes in cancer incidence.

Table 5: Lung cancer, subboundary statistics.

Boundaries	Statistic	Observed	Expected	P↑	P↓
Female lung cancer	Ns	40	44	0.800	0.260
	Lmean	2.450	2.254	0.260	0.800
	Lmax	11.000	10.080	0.384	0.712
Male lung cancer	Ns	43	44	0.648	0.456
	Lmean	2.279	2.231	0.456	0.648
	Lmax	9.000	9.442	0.576	0.588
OPR for lung cancer	Ns	143	204	1.000	0.004
	Lmean	4.783	3.261	0.004	1.000
	Lmax	110.00	18.727	0.004	1.000

cially in the central portion of Long Island, but this amount of overlap was not statistically significant under boundary overlap analysis when Long Island is considered as a whole (Table 6). This indicates that the geographic distributions of male and female lung cancer SMR differ and are possessed of boundaries whose placement appears independent. This implies the geographic determinants of lung cancer differ for males and females.

Lung Cancer – Analysis of NATA Data

The overall predicted risk (OPR) for lung cancer was calculated from the NATA dataset and mapped (Figure 14). We see a broad area of higher overall predicted risk in the

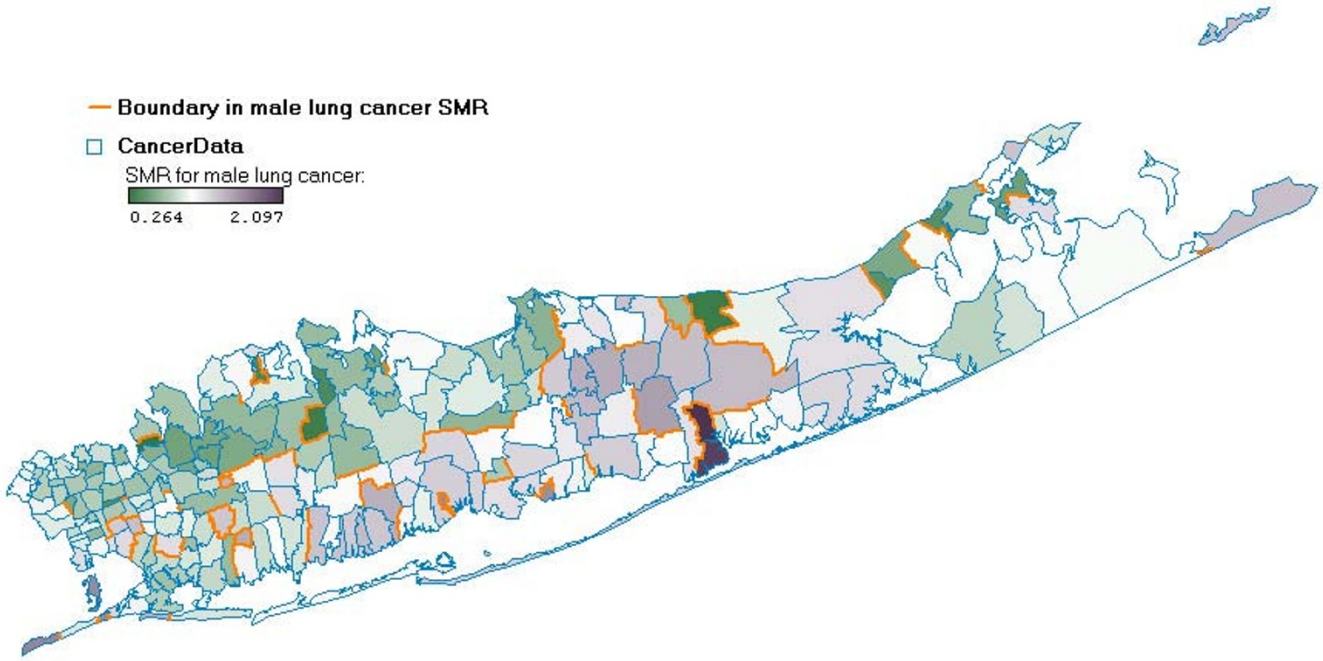

Figure 11
Male lung cancer incidence showing local boundaries (orange). The blue outlines are the ZIP code edges. The fill in the ZIP code areas indicates the SMR for male lung cancer, with darker purple regions having higher SMR, white regions having SMR near 1, and darker green regions having lower SMR. The boundaries shown in orange indicate those ZIP code edges with large changes in cancer incidence.

central section of Long Island on an axis from the vicinity of Northport and Saint James south to East Islip and Babylon. A smaller, isolated area of elevated risk is found near Flushing, East Elmhurst (see map inset). The corresponding ZIP code (11369) is included in the local Moran cluster for significantly low female lung cancer SMR [1]. Hence this spatial outlier in high lung cancer OPR is not predictive of high lung cancer incidence in females.

The large clusters in central Long Island for both male and female lung cancer SMR [1] at least partially overlap the area of high and moderately high OPR in central Long Island, with some offset of the lung cancer clusters to the East. Is this apparent overlap statistically significant? We first undertake the subboundary analysis of the NATA data, and then assess boundary overlap between OPR and lung cancer.

Subboundary Analysis

Figure 14 shows OPR for lung cancer, overlaid with boundaries in lung cancer OPR. The boundaries in lung cancer OPR are significantly long and contiguous (Table 5). There are fewer boundaries composed of only one boundary element than is expected (Ns $P\downarrow = 0.004$) under the null hypothesis of no spatial pattern in lung cancer OPR values. The boundaries in OPR also have a longer mean and maximum length than expected (Lmean $P\uparrow = 0.004$; Lmax $P\uparrow = 0.004$). This indicates that zones of rapid change in OPR values are long and contiguous. This result is consistent with the map of OPR that shows a broad area of elevated risk in central Long Island.

Overlap Analysis

If the air toxics modeled in the NATA database were indeed associated with lung cancer, then we would expect boundaries in OPR to overlap boundaries in lung cancer (Figure 15). We found statistically significant overlap between boundaries in female lung cancer and OPR (Table 6). The average minimum distance from a boundary in female SMR to the nearest boundary in OPR is significantly smaller (Og $P\downarrow = 0.044$) than the average minimum distance expected under a null hypothesis of no spatial pattern in both cancer incidence and overall predicted risk. While boundaries in female lung cancer SMR were significantly close to boundaries in OPR, boundaries in OPR are further from boundaries in female lung cancer incidence than is expected by chance (Oh $P\uparrow = 0.002$). This result indicates that while boundaries in female lung can-

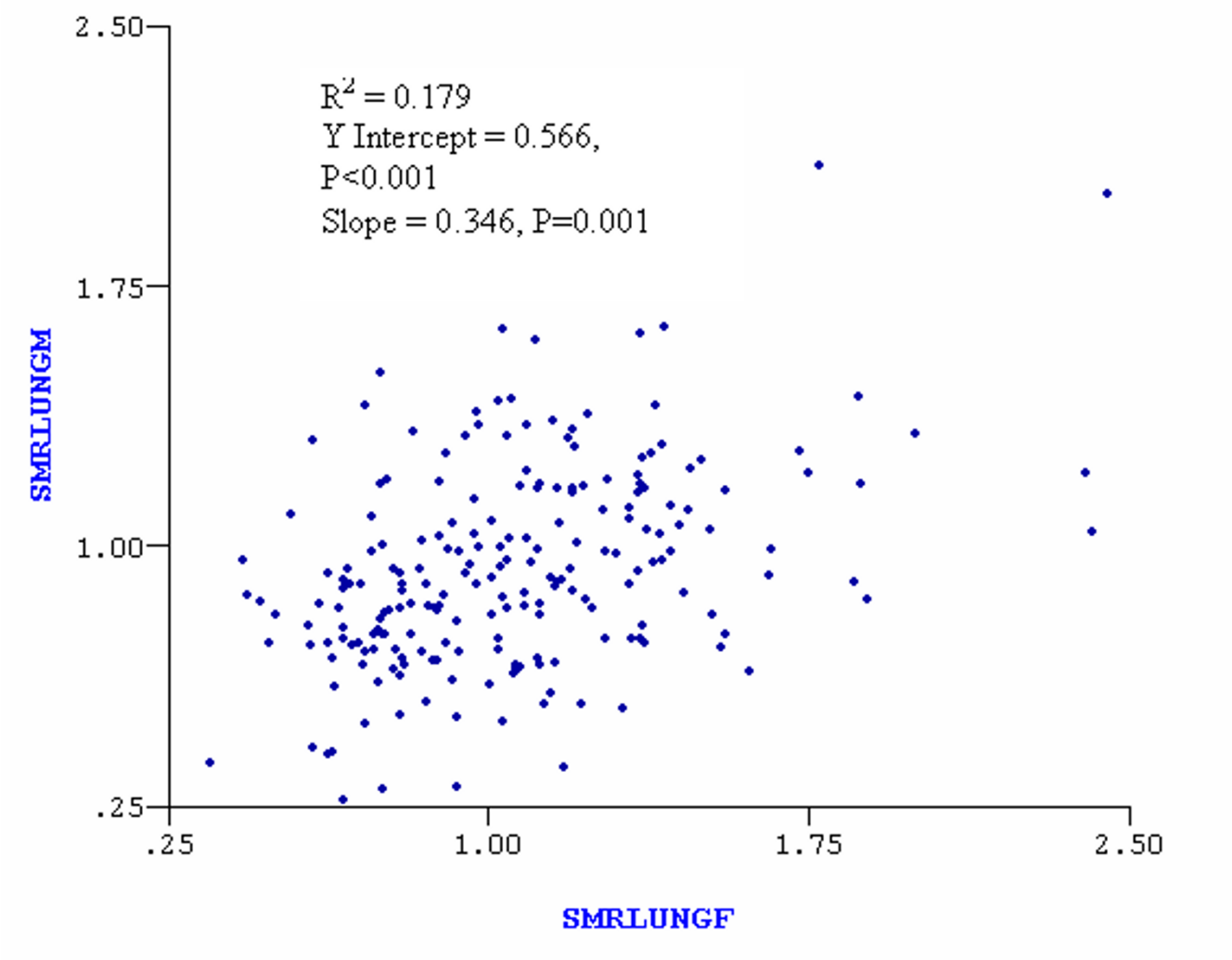

Figure 12
Scatter gram of male lung cancer versus female lung cancer SMR. There is a weak positive association between male and female lung cancer incidence. Male lung cancer SMR has a higher maximum value (near 2.5 SMR) than the maximum female value (near 2.0 SMR), perhaps reflecting greater smoking prevalence among males.

cer SMR are most likely to occur near boundaries in the NATA air toxics, boundaries in the NATA air toxics do not always occur near boundaries in female cancer incidence. This pattern of results is consistent with a causal model where female lung cancer depends on the overall predicted risk from the air pollutants modeled in the NATA dataset. However, it is not a demonstration of causality.

A similar result obtains for males (Table 6), where the average minimum distance from a boundary in male lung cancer SMR to the nearest boundary in lung cancer OPR is significantly smaller than its expected value (Og P↓ = 0.044). While boundaries in male lung cancer incidence are significantly near boundaries in OPR, the boundaries in OPR are significantly farther from boundaries in male lung cancer incidence than is expected by chance (Oh P↑ = 0.009). As for female lung cancer, this is consistent with a causal model where male lung cancer incidence depends on the overall predicted risk as estimated by the NATA database.

Discussion

To summarize, this geographic study found little or no association between the incidence of colorectal and breast cancers and the overall predicted risks from air toxics as modeled by the National Air Toxics Assessment Program.

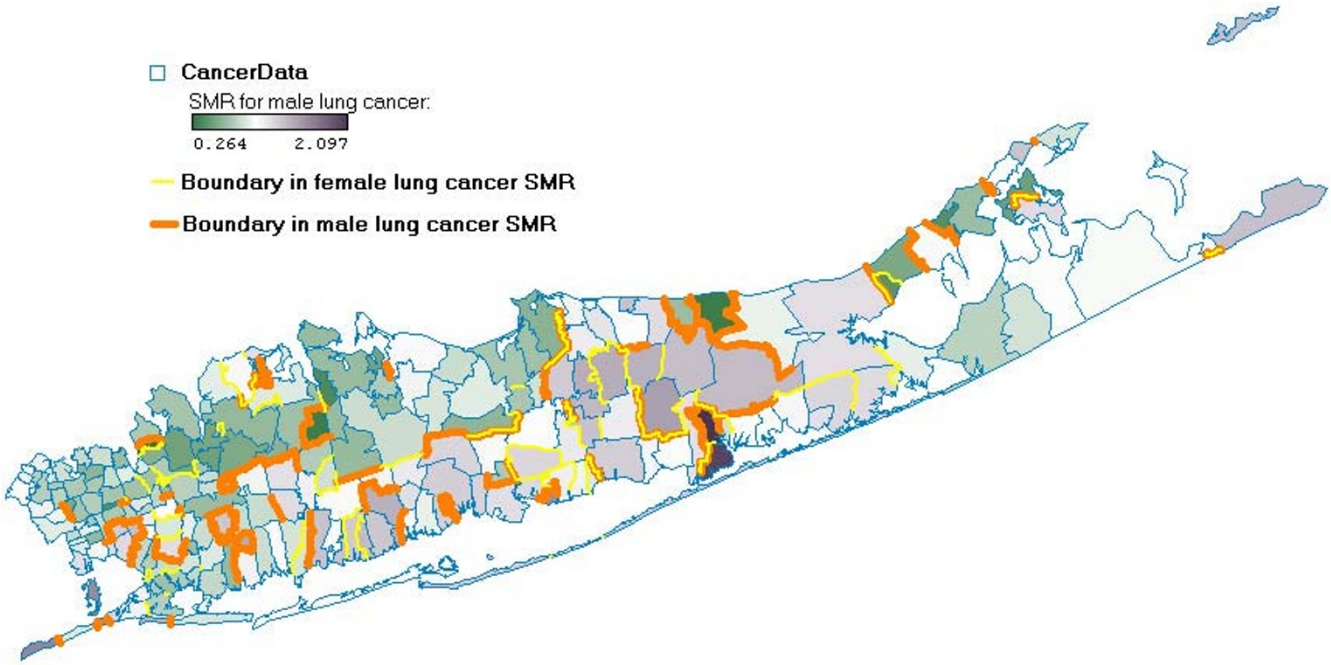

Figure 13
Map of male and female lung cancer incidence and boundaries. This map shows the boundaries for male and female lung cancer superimposed on the map of male lung SMR. The blue lines are ZIP code edges. The boundaries shown as thin yellow lines indicate those ZIP codes with large changes in female lung cancer incidence. The boundaries shown in thick orange lines indicate the edges of ZIP codes across which there are large changes in male lung cancer incidence. Locations where the thick orange line surrounds the yellow line illustrate exact boundary overlap.

Table 6: Lung cancer, overlap statistics.

Boundaries overlapped	Statistic	Observed (meters)	Expected (meters)	P↑	P↓
Male (g) and female (h) lung cancer boundaries	Og	2436.668	1644.780	0.052	0.952
	Oh	1353.433	1686.974	0.752	0.252
	Ogh	1895.050	1665.829	0.236	0.768
	Os	48	48.671	0.580	0.484
Female lung cancer boundaries (g) and OPR boundaries (h)	Og	495.163	908.754	0.958	0.044
	Oh	5085.973	1804.606	0.002	1.000
	Ogh	4503.956	1678.020	0.002	1.000
	Os	0.000	0.136	1.000	0.872
Male lung cancer boundaries (g) and OPR boundaries (h)	Og	499.713	886.212	0.957	0.044
	Oh	2573.460	1803.578	0.009	0.992
	Ogh	2312.912	1685.834	0.013	0.988
	Os	0.000	0.139	1.000	0.874

Table 7: UREs for air toxics

Compound	Associated with Cancers			URE (excess cases per $\mu g/m^3$)
	Breast	Colorectal	Lung	
1,3-Butadiene	x			1.0×10^{-5}
1,3-Dichloropropene			x	4.0×10^{-6}
Acrylonitrile	x		x	6.8×10^{-5}
Arsenic			x	4.3×10^{-3}
Benzene	x			7.8×10^{-6}
Beryllium			x	2.4×10^{-3}
Cadmium			x	1.8×10^{-3}
Carbon Tetrachloride	x			1.5×10^{-5}
Chloroform		x		2.3×10^{-5}
Chromium			x	1.2×10^{-2}
Diesel Particulate Matter			x	8.15×10^{-4}
Ethylene Dibromide	x		x	2.2×10^{-4}
Ethylene Dichloride	x		x	2.6×10^{-5}
Ethylene Oxide	x		x	8.8×10^{-5}
Hydrazine			x	4.9×10^{-3}
Methylene Chloride	x			4.7×10^{-7}
Nickel	x		x	1.2×10^{-4}
Polycyclic Aromatic Hydrocarbons	x		x	2.0×10^{-4}
Perchloroethylene		x	x	5.9×10^{-6}
Trichloroethylene			x	2.0×10^{-6}
Vinyl Chloride	x		x	8.8×10^{-6}

This study did find association between geographic patterns in incidence of male and female lung cancer and the overall predicted risks for lung cancer. This association was found at two levels by different statistical methods. First, clusters of statistically significant high and low lung cancer incidences were identified separately for both males and females [1]. Both males and females demonstrated a large cluster of high lung cancer incidence in central Long Island that corresponded approximately, in both geographic extent and location, to a broad zone of high overall lung cancer risk as predicted by the air toxics estimates. Second, the appearance of an association on maps is of course subjective when only the human eye is employed, and we further assessed statistical validity of map pattern using geographic boundary analysis. Boundary analysis found that boundaries in both male and female lung cancer incidence are significantly near to boundaries in overall predicted risk, and that, while boundaries in lung cancer SMR tend to always be near boundaries in OPR, boundaries in OPR are not necessarily near to boundaries in cancer SMR. This is consistent with a causal model where geographic pattern in lung cancer is determined, at least in part, by geographic pattern in the overall predicted risk due to air toxics.

Demonstration of a causal relationship between exposure to the modeled air toxics and lung cancer incidence clearly is beyond the inferential ability of this study. While consistent with a causal relationship between air toxics and lung cancer, these results do not demonstrate the existence of a causal relationship. In fact, the NATA air toxics dataset models 1996 exposure, while the cancer dataset covers the period from 1993–7. We cannot argue that 1996 modeled exposure caused 1993 cancers. However, if the NATA 1996 model is consistent with geographic variation in exposures to air toxics in the decade before 1993 (a time period that includes the latency of these cancers), a causal relationship cannot be precluded. This is an opportunistic analysis, taking advantage of existing publicly available historical data, and for which the ideal data do not exist. Now that geographic variation has been quantified, additional explanatory variables such as smoking incidence, socioeconomic status, education, occupation and ethnicity (each of which is known to be an important determinant of lung cancer incidence and/or smoking behavior) can be incorporated into the analysis. We also note that examination of spatial outliers in OPR for lung cancer found that these outliers were not necessarily associated with elevated lung cancer risk.

The results from this study can be used to guide further investigation, especially to identify areas where the associations between boundaries in OPR and in lung cancer incidence are strongest. By map inspection (refer to Figure

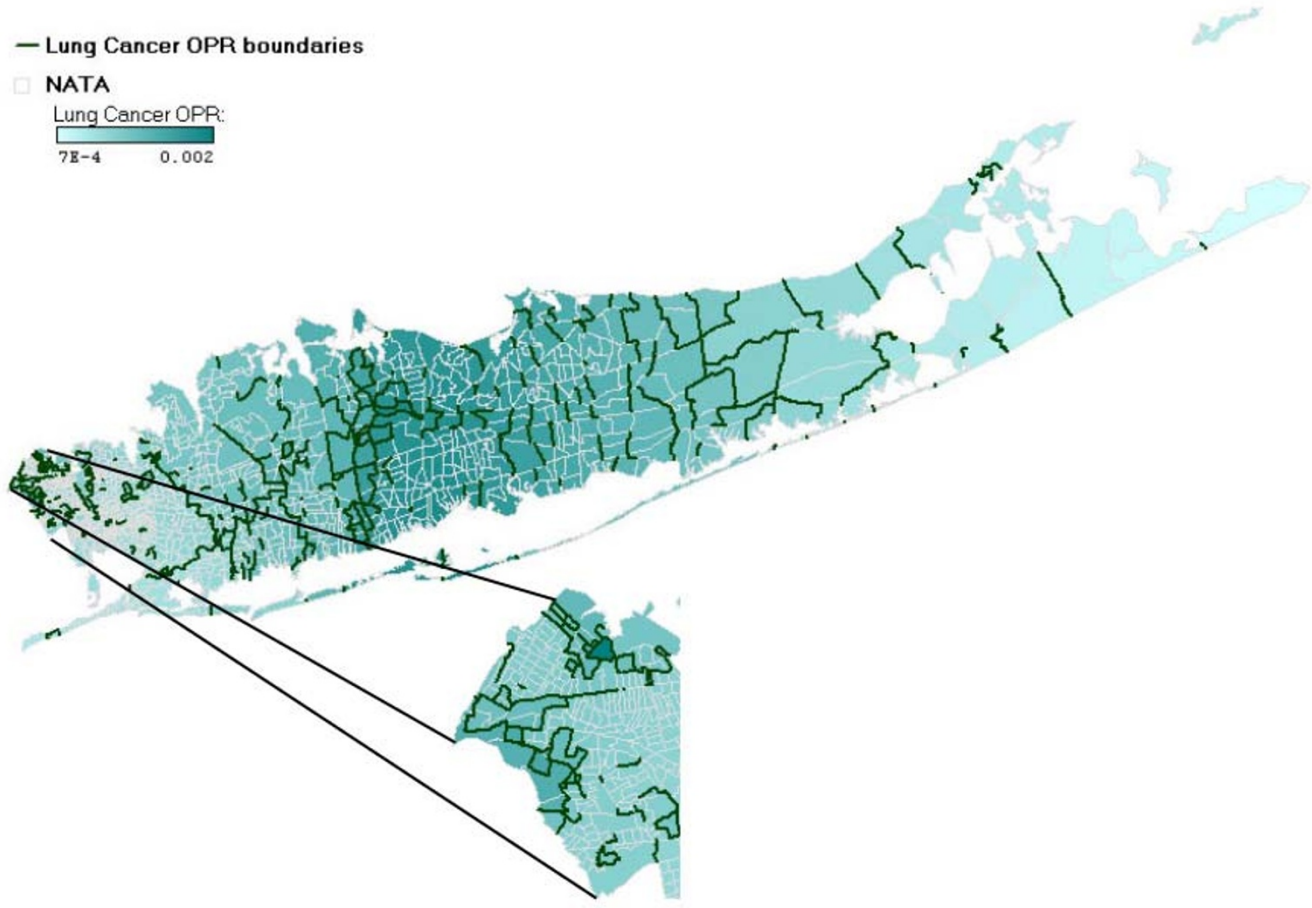

Figure 14
Map of overall predicted risk (OPR) for lung cancer with boundaries. The turquoise fill gradient indicates the OPR for lung cancer, with darker regions having higher OPR, and the gray lines define the census tract edges. Dark green lines represent boundaries in OPR for lung cancer.

15) these areas of strong boundary overlap include the vicinities of Locust Valley (11560) and Mill Neck (11765); Bohemia (11716) and Islip Terrace (11752); and Mastic Beach (11951) and Shirley (11967). Each of these is a "zone of rapid change" where both cancer morbidity and OPR change dramatically over a relatively short distance. By focusing future inquiry on locations where boundaries in both cancer morbidity and the putative exposure overlap, we should be able to better identify the local- as well as large-scale determinants of lung cancer.

Conclusions

In general, geographic studies of encountered data, such as this one, have power only to reject, and not confirm, predictions founded on scientific hypotheses. For example, one hypothesis implicit in our study is: "Lung cancer is at least partly determined by exposure to the relevant air toxics modeled in the NATA program." And the corresponding prediction is "Geographic variation in lung cancer is at least partly determined by geographic variation in the NATA overall predicted risk." While we may now conclude, within the framework of the methods employed, that geographic variation in lung cancer is indeed associated with geographic variation in the NATA overall predicted risk for lung cancer, we cannot conclude that exposure to modeled air toxics caused lung cancer in the study population. The existence of a geographic association is not sufficient to demonstrate causality, particularly given the latency of cancer and the time span of the available environmental data.

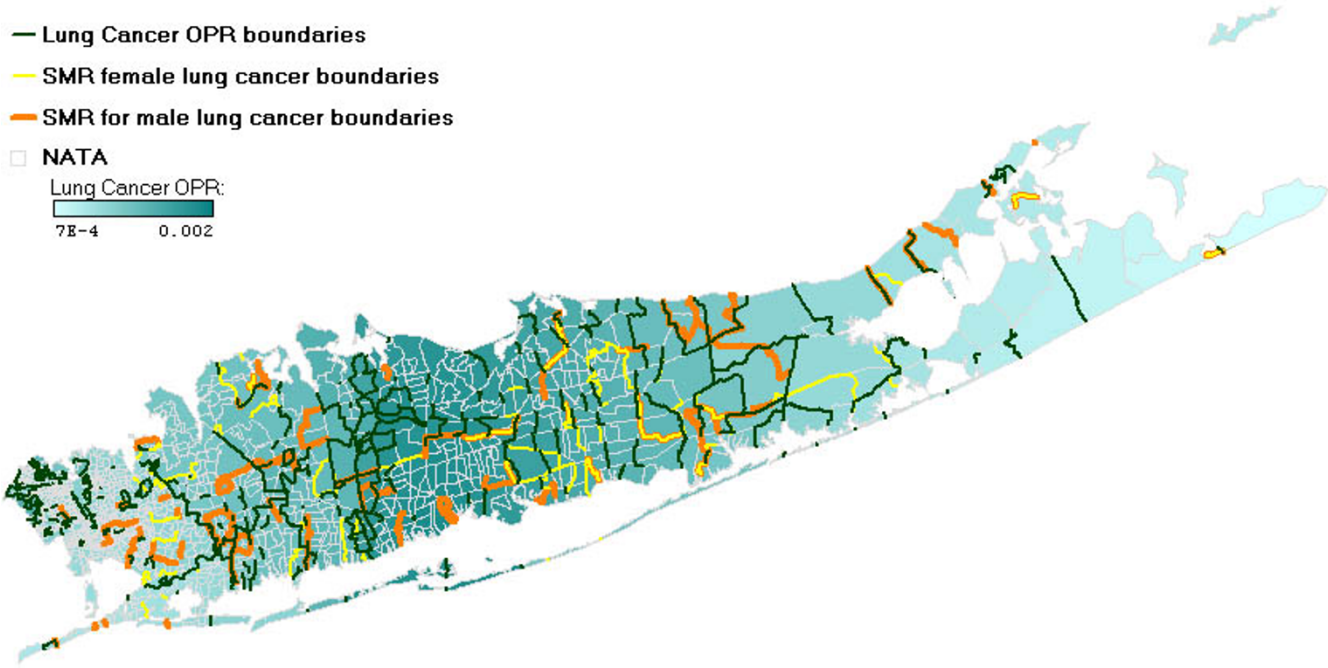

Figure 15
Map of lung cancer OPR and SMR boundaries. The turquoise fill gradient indicates the OPR for lung cancer, with darker regions having higher OPR, and the gray lines define the census tract edges. Boundaries in lung cancer OPR are indicated in dark green, along with boundaries in male (orange) and female (yellow) lung cancer incidence. Boundaries in male and female lung cancer SMR do not overlap with one another to a significant extent. However, boundaries in both male and female lung cancer incidence do significantly overlap with OPR boundaries.

Ecologic fallacy, location uncertainty and the use of residence as an exposure surrogate

This study demonstrated the existence of statistically significant clusters of both cancer excess and deficit [1]. The evaluation of geographic clusters must include consideration of potentially misleading aspects of ecologic studies. ZIP codes and census tracts are a coarse spatial unit for aggregating cancer cases and for estimating individual exposure to air toxics. The ZIP codes and census tracts also are measured at different spatial scales, although this is controlled for in the statistical tests by employing randomization procedures based on ZIP code and census tract geography. In the long term, other spatial divisions may be more appropriate for estimating environmental exposure – watersheds, aquifers, or local public water systems for water-borne substances, or "windsheds" for airborne substances. Yet, because of privacy concerns for the patients and the limitations on existing environmental data, we used the available data.

Also, the ZIP code of residence at diagnosis is an inadequate descriptor of an individual's location during the development of cancer. Using the ZIP code of residence assumes the patient lived within that ZIP code area during the period of time required to develop cancer following exposure to an environmental compound that influenced cancer risk. Hence the degree of exposure to the potential risk factors over a multi-year period has been estimated for each study subject based on their place of residence, aggregated at the census tract level. This assumption is clearly tenuous given the mobility of the study population. Additionally, the patient may have worked and spent a great deal of time in another area. The use of residence as a surrogate for exposure clearly is invalid whenever causative exposures largely occur outside of the home – for this study "outside of the home" means in a different ZIP code zone.

Confounders and covariates

This study accounted for age but ignored other confounders and covariates such as socioeconomic status, occupation, risk behaviors (such as smoking and diet) and ethnicity. No attempt was made to account for genetic predispositions to cancer that are known to be associated with a person's heritage. Thus the spatial patterns we ob-

served in the cancer SMR's could be caused by geographic variation in these confounders and covariates.

Cancer latency and migration

The cancers explored are known to have long latencies on the order of 10 to 20 years, depending on the causative exposure and the subject's genetic predisposition. Over such a time span the average American moves several times, tending to obscure whatever geographic relationships may exist between environmental exposure and cancer incidence.

Higher-order interactions

We used boundary overlap analysis to assess potential bivariate association between risk estimates and cancer incidence. Especially for complex relationships (such as those between environment and cancer), apparent bivariate associations may be driven by multivariate interactions that are not directly quantified by the two variables under scrutiny. For example, elevated air pollution may lead to lower housing prices, which in turn attracts poorer households with higher smoking rates. In this instance, an observed bivariate correlation between air pollution and cancer would not be indicative of the underlying causal mechanism.

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 Allen, environmental epidemiologist and former US EPA program director for the Long 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 **Local clustering in breast, lung, and colorectal cancer on Long Island, NY.** *Int J Health Geogr* 2003, **2:**3
2. Susser M and Susser E **Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology.** *American Journal of Public Health* 1996, **86:**674-7
3. United States Environmental Protection Agency **DRAFT National-Scale Air Toxics Assessment for 1996.** *Office of Air Quality Planning And Standards. EPA-453/R-01-003* 2001, [http://www.epa.gov/ttn/atw/sab/natareport.pdf]
4. California Environmental Protection Agency **Air toxics hot spots program risk assessment guidelines. Part II: Technical support document for describing available cancer potency factors.** *Office of Environmental Health Hazard Assessment, Air Toxicology and Epidemiology Section* 1999, [http://www.oehha.ca.gov/pdf/HSCA2.pdf]
5. Womble WH **Differential systematics.** *Science* 1951, **114:**315-322
6. Barbujani G, Oden NL and Sokal RR **Detecting areas of abrupt change in maps of biological variables.** *Systematic Zoology* 1989, **38:**376-389
7. Bocquet-Appel JP and Bacro JN **Generalized wombling.** *Systematic Zoology* 1994, **43:**442-448
8. Fortin M-J **Effects of data types on vegetation boundary delineation.** *Canadian Journal of Forest Research* 1997, **27:**1851-1858
9. Fortin M-J **Edge detection algorithms for two-dimensional ecological data.** *Ecology* 1994, **75:**956-965
10. Fortin M-J and Drapeau P **Delineation of ecological boundaries: Comparisons of approaches and significance tests.** *Oikos* 1995, **72:**323-332
11. Oden NL, Sokal RR, Fortin M-J and Goebl H **Categorical wombling: Detecting regions of significant change in spatially located categorical variables.** *Geographical Analysis* 1993, **25:**315-336
12. Jacquez GM **The map comparison problem: Tests for the overlap of geographic boundaries.** *Stat Med* 1995, **14:**2343-2361