

Methodology

Open Access

## In search of induction and latency periods: Space-time interaction accounting for residential mobility, risk factors and covariates

Geoffrey M Jacquez\*<sup>1,2</sup>, Jaymie Meliker<sup>1</sup> and Andy Kaufmann<sup>1</sup>

Address: <sup>1</sup>BioMedware, Ann Arbor, USA and <sup>2</sup>Department of Environmental Health Sciences, The University of Michigan, Ann Arbor, USA

Email: Geoffrey M Jacquez\* - jacquez@biomedware.com; Jaymie Meliker - meliker@biomedware.com;  
Andy Kaufmann - afsb@biomedware.com

\* Corresponding author

Published: 23 August 2007

Received: 30 May 2007

*International Journal of Health Geographics* 2007, **6**:35 doi:10.1186/1476-072X-6-35

Accepted: 23 August 2007

This article is available from: <http://www.ij-healthgeographics.com/content/6/1/35>

© 2007 Jacquez et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Space-time interaction arises when nearby cases occur at about the same time, and may be attributable to an infectious etiology or from exposures that cause a geographically localized increase in risk. But available techniques for detecting interaction do not account for residential mobility, nor do they evaluate sensitivity to induction and latency periods. This is an important problem for cancer, where latencies of a decade or more occur.

**Methods:** New case-only clustering techniques are developed that account for residential mobility, latency and induction periods, relevant covariates (such as age) and risk factors (such as smoking). The statistical behavior of the methods is evaluated using simulated data to assess type I error (false positives) and statistical power. These methods are applied to 374 cases from an ongoing study of bladder cancer in 11 counties in southeastern Michigan, and the ability of the methods to localize space-time interaction at the individual-level is demonstrated.

**Results:** Significant interaction is found for induction periods of ~5 years and latency ~19.5 years. Data are still being collected and the observed clusters may be attributable to differential sampling in the study area.

**Conclusion:** Residential histories are increasingly available, raising the possibility of routine surveillance in a manner that accounts for individual mobility and that incorporates models of cancer latency and induction. These new techniques provide a mechanism for identifying those geographic locations and times associated with increases in cancer risk *above and beyond* that expected given covariates and risk factors in geographically mobile populations.

### Background

Cluster analysis provides an objective basis for evaluating whether geographic cancer patterns are significant [1,2]. Dozens of approaches are now available (e.g., [3-10]); however, most of these were developed for spatially static datasets and assume individuals are immobile and that latency is negligible [11]. Most published studies still rely

only on place of residence at time of diagnosis or of death to record the locations of health events. But when analyzing cancers, causative exposures may occur many years prior to diagnosis, and during this interval individuals may move place of residence. Failure to account for residential mobility, therefore, can make detecting clustering of cases in relation to causative exposures difficult or even

impossible. Recent studies demonstrate that results obtained using static spatial point distributions can lead to erroneous conclusions regarding the timing, existence, extent, and locations of disease clusters [12,13]. Tests for space-time interaction that account for residential mobility thus are required when studying cancer.

For cancer, interaction statistics allow researchers to explore two different types of etiological hypotheses: infectious processes (e.g. cancers with viral origins), and geographically and temporally localized exposures to carcinogenic agents (e.g. exposure to radon in home environments). In addition, interaction tests have the substantial advantage of working with cases-only data, and do not require the selection of controls. The development of appropriate interaction tests that account for residential mobility, risk factors, covariates and reasonable models of latency and induction periods is expected to be a significant methodological advance that will allow researchers to work directly with data from cancer registries without the need for the painstaking selection of matched controls.

In 1967 Nathan Mantel [14] proposed a space-time interaction test for case data, and represented the observations as  $\{x_i, y_i, t_i\}$ . Here  $x_i, y_i$  is the place of residence for the  $i^{\text{th}}$  case, and  $t_i$  is the time of diagnosis or death. "Interaction" arises when nearby cases occur at about the same time, and may indicate a contagious process such as infection transmission, or a geographically and temporally localized exposure to a carcinogen. For infection the underlying assumption is that nearby individuals are more likely to interact and experience infection transmission events. For a localized exposure the assumption is that nearby individuals will experience similar exposures such that their disease risk will be elevated at about the same time.

The proximity metrics underlying Mantel's test are the spatial and temporal distances between pairs of cases. Knox [15] used adjacencies, Diggle et al. [16] the K-function and Jacquez [17] nearest neighbor relationships. Recent adaptations to Knox's method account for changing population size [18] and the time required for infection transmission [19], but do not account for human mobility. In studies of cancer clustering, methods have yet to effectively account for latency, perhaps because latency is difficult to observe, and our knowledge of it is uncertain. This becomes increasingly problematic when we consider residential mobility. The average American now moves every 5-7 years, meaning that at time of diagnosis few cases actually reside where causative exposures may have occurred [20]. And no tests for interaction simultaneously account for human mobility, latency, risk factors and covariates. This paper introduces novel techniques that account for residential mobility, cancer latency, risk

factors and covariates, evaluates them using simulations, and then applies them in a study of bladder cancer in southeastern Michigan.

**Methods**

We begin with descriptions of the empirical induction period (EIP), notation, models of EIP and metrics for evaluating residential proximity for mobile individuals. We then derive space-time interaction tests that incorporate EIP and residential mobility. Next, we extend these to adjust for risk factors and covariates. We then define the algorithm used to evaluate sensitivity of the interaction statistics to specification of the EIP. Finally, we apply the new methods to (a) simulated data for which the extent of interaction is known and (b) residential histories of bladder cancer cases in Michigan.

Rothman [21] recognized that illness in an individual may have a multiplicity of causes, none of which alone may be sufficient to cause the disease. This makes definition and observation of disease latency problematic. He recommended that one explore sensitivity of latency-based metrics by evaluating a range of plausible empirical induction periods. We define the EIP as an induction period,  $\omega$ , in which causative exposures occurred, and a lag,  $\tau$ , the latency. In practice  $\omega$  and  $\tau$  are unobservable, and we therefore explore sensitivity of interaction to specification of these parameters.

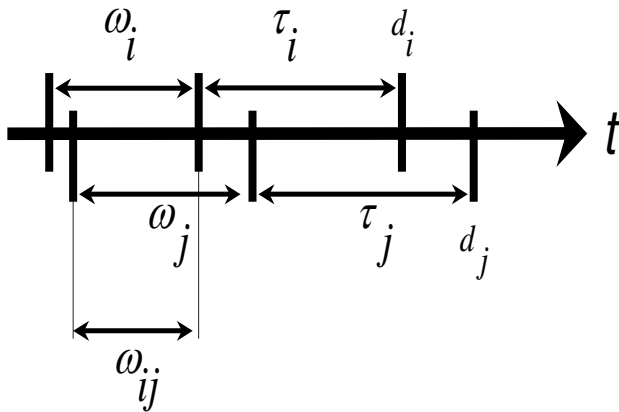
Let  $d_i$  represent the time of diagnosis of case  $i$ . This could be time of death or another event in the life course, but for exposition we use time of diagnosis. The locations where a person resides during  $\omega$  is called the *exposure trace* [12]. We subscript the induction period,  $\omega_i$ , and latency,  $\tau_i$ , so that they can differ across cases. Now consider cases  $i$  and  $j$ . Define  $\omega_{ij}$  as the interval when  $\omega_i$  overlaps  $\omega_j$  (Figure 1, Equation 1).

$$\omega_{ij} = \omega_i \cap \omega_j \quad (1)$$

A measure that accounts for residential mobility and co-occurrence of induction periods is then

$$\eta_{ijk\omega} = \begin{cases} 1 & \text{iff } i \text{ and } j \text{ were ever } k \text{ nearest neighbors during } \omega_{ij} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

It is 1 if the places of residence of cases  $i$  and  $j$  were ever  $k$ -nearest neighbors during  $\omega_{ij}$ . Hence  $\eta_{ijk\omega}$  is 1 if cases  $i$  and  $j$  lived near one another at some time when their induction periods overlapped. If their induction periods never overlapped or if they were not  $k$  nearest neighbors then  $\eta_{ijk\omega}$  is zero.



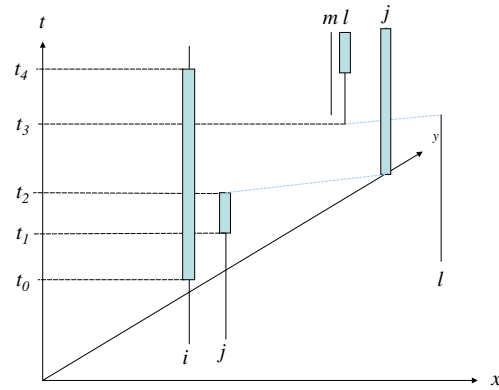
**Figure 1**  
Model of empirical induction periods. The date of diagnosis for the  $i^{\text{th}}$  case is  $d_i$ .  $\tau_i$  is the temporal lag between initiation of the disease (e.g. appearance of the first cancer cells) and diagnosis.  $\omega_i$  is the induction period when causative exposures occurred.  $\omega_j$  is that time interval when the induction windows for cases  $i$  and  $j$ ,  $\omega_i$  and  $\omega_j$ , overlapped.

**Local test accounting for residential mobility and EIP**  
Let  $N$  be the total number of cases. A local statistic for mobile individuals that accounts for the induction period is

$$V_{ik\omega} = \sum_{\substack{j=1 \\ j \neq i}}^N \eta_{ijk\omega}. \tag{3}$$

We call this the local Vesta statistic after the Roman Goddess of the hearth. It is the count of the  $k$ -nearest neighbors of case  $i$  whose induction periods overlapped those of case  $i$ . This statistic is evaluated about the residential history for each case, and assesses whether and where there is interaction about that case's exposure trace. Its statistical significance is assessed by holding the residential histories constant, and by randomizing the dates of diagnosis with equal probability across the residential histories. The null hypothesis is that an observed date of diagnosis is equiprobable across the  $N$  cases.

It is possible for  $V_{ik\omega}$  to exceed  $k$ , since the geometry of the residential histories changes through time and  $V_{ik\omega}$  is incremented over case  $i$ 's exposure trace. To illustrate in Figure 2  $x$  and  $y$  indicate geographic space and the vertical axis is time. The residential histories for case  $i$ ,  $j$ , and  $l$  are shown as vertical lines. Case  $i$  never moves and is shown as a continuous, vertical line through time. Exposure traces are shown by long rectangles about a residential his-



**Figure 2**  
Dynamic topology of residential histories and exposure traces. See text.

tory. For example,  $\omega_i$  is indicated by the rectangle about the residential history for subject  $i$  from  $t_0$  to  $t_4$ . Notice case  $l$  moved place of residence at  $t_3$ , and that case  $j$  moved at  $t_2$  during its induction period  $\omega_j$ . Using  $k = 1$  nearest neighbors we see that:

$V_{i1\omega} = 1$ , since  $\eta_{ji1\omega} = 1$  from  $t_1$  to  $t_2$  when  $i$  and  $j$  were 1<sup>st</sup> nearest neighbors.

$V_{j1\omega} = 2$ , since  $\eta_{ji1\omega} = 1$  from  $t_1$  to  $t_2$  when  $i$  was the 1<sup>st</sup> nearest neighbor of  $j$ , and  $\eta_{li1\omega} = 1$  from  $t_3$  to  $t_4$  when  $l$  was the 1<sup>st</sup> nearest neighbor of  $j$ .

$V_{l1\omega} = 0$ , since case  $m$ , the first nearest neighbor to  $l$ , did not have an active exposure trace and  $\eta_{ml1\omega} = 0$ .

$V_{m1\omega} = 0$ , since case  $m$ 's exposure trace never overlapped any others.

**Duration-weighted local interaction statistic**

We can extend this to account for the duration of residential stays. Define the duration of time when the induction periods for  $i$  and  $j$  overlapped and when  $j$  was a  $k$  nearest neighbor of case  $i$ , and write it as  $\Delta\eta_{ijk\omega}$ . A duration weighted local Vesta is

$$\Delta V_{ik\omega} = \sum_{\substack{j=1 \\ i \neq j}}^N \Delta\eta_{ijk\omega}. \tag{4}$$

The units on this statistic are person time (e.g. case days). It quantifies the number of days during case  $i$ 's induction period when its  $k$ -nearest neighbors were also in their induction periods. Suppose  $\Delta V_{i2\omega} = 2$ . This means the induction period for one of its  $k = 2$  nearest neighbors was

"active" for 2 days during case  $i$ 's induction period, or that both of its  $k = 2$  nearest neighbors had active induction periods of 1 day during case  $i$ 's induction period.

**Risk factor and covariate adjustment**

We may have knowledge of risk factors and covariates as when a case-control study has been conducted on a subset of the available data. One then can quantify the probability of a given participant being a case, given the risk factors and covariates [22]. Let  $p_i$  denote the probability of participant  $i$  being a case given their vector of risk factors and covariates  $x_i$ . We would like to construct a version of the local statistic that is sensitive to interaction *above and beyond* that attributable to geographic variation in known risk factors and covariates. We accomplish this by giving decreased weight to those individuals whose cancers are likely attributable to the risk factors and covariates, allowing us to focus our attention on interaction in those cases whose etiology is largely unexplained. For the local Vesta adjusted for covariates

$$V_{ik\omega x} = (1 - p_i) \sum_{\substack{j=1 \\ i \neq j}}^N \eta_{ijk\omega} (1 - p_j), \tag{5}$$

and

$$\Delta V_{ik\omega x} = (1 - p_i) \sum_{\substack{j=1 \\ i \neq j}}^N \Delta \eta_{ijk\omega} (1 - p_j) \tag{6}$$

for the duration-weighted version. Here  $p_i$  denotes the probability of participant  $i$  being a case given their vector of risk factors and covariates  $x_i$ . Hence the terms  $(1 - p_j)$  and  $(1 - p_i)$  effectively discount the contributions of cases  $j$  and  $i$  (respectively) when their cancers reasonably might be attributable by known risk factors and covariates. In practice one will want to calculate the statistics twice, the first time using Equation 4, and the second time adjusting for risk factors and covariates using Equation 6. Comparison of the results identifies cases for which space-time interaction is explained by the risk factors and covariates, and those that are significant both before and after statistical adjustment.

**Global interaction statistics**

Equations 3 and 4 quantify local interaction about specific cases. Global tests that assess interaction when all of the cases are considered simultaneously are

$$V_{k\omega} = \sum_{i=1}^N V_{ik\omega} \tag{7}$$

and

$$\Delta V_{k\omega} = \sum_{i=1}^N \Delta V_{ik\omega}. \tag{8}$$

Equation 7 is an integer count and Equation 8 is duration-weighted. In practice the duration-weighted version is preferred since the duration when exposure traces overlap is of epidemiological interest. When information regarding the probability of being a case is available the global statistics are

$$V_{k\omega x} = \sum_{i=1}^N V_{ik\omega x} \tag{9}$$

and

$$\Delta V_{k\omega x} = \sum_{i=1}^N \Delta V_{ik\omega x}. \tag{10}$$

Here the subscript  $k\omega x$  denote the number of  $k$  nearest neighbors being considered ( $k$ ), the induction period ( $\omega$ ) and the vector of covariates and risk factors  $x$  for that case.

**Local spatial clustering of exposure traces at time  $t$**

Equations 3–6 are accumulated over the exposure traces in the individual life histories. We calculate these local statistics through time, then inspect time plots for shape and inflection points on these monotonically increasing step functions. But because the local Vesta statistics are accumulated over time, they are not particularly sensitive to an ephemeral clustering of exposure traces, since the "signal" added by such clustering is diluted by all that has gone before. We therefore desire a test for local spatial clustering of exposure traces at any given time  $t$ . We would like this statistic to tell us, when considering case  $i$ , whether its  $k$ -nearest neighbors tend to have "active" exposure traces. Define

$$c_{it} = \begin{cases} \text{IFF case } i \text{ is in its exposure trace at time } t \ (t \in \omega_i) \\ 0 \text{ otherwise} \end{cases} \tag{11}$$

The spatial clustering test is then

$$S_{ik\omega t} = c_{it} \sum_{j=1}^k c_{jt} \tag{12}$$

The summation is over case  $i$ 's  $k$  nearest neighbors. We call this the Janus statistic, after the Roman God who guarded the doorway to the home. Janus is the count, at time  $t$ , of the number of  $k$  nearest neighbors of case  $i$  with overlapping induction periods. Notice the statistic can be non-zero only when case  $i$  is in its induction period. If we define the time interval  $\Delta t$  such that the geography of the residential histories doesn't change (e.g. none of the cases

moves place of residence, and whether case  $i$  and its neighbors are in their respective induction periods doesn't change) we may consider the time weighted version of the statistic

$$\Delta S_{ik\omega\tau} = \Delta t \sum_{j=1}^k c_{jt} \quad (13)$$

This statistic is measured in case-time units, e.g. case-days.

#### **Focused spatial clustering of exposure traces at time $t$**

Suppose we know the address history of a putative source of a carcinogen, such as an industry. Given focus  $f$  we denote this address history as  $F_f$ . Further suppose we have information regarding the emission volume per unit time, such as might come from EPA's TRI (Toxic Release Inventory) data. Call this  $E_f(t)$ . The  $i^{\text{th}}$  case has induction period  $\omega_i$  that begins at  $t_{i0}$  and ends at  $t_{i1}$ . An emission-weighted focused Vesta statistic is then

$$\Delta V_{fik\omega} = \sum_{i=1}^k \int_{t_{i0}}^{t_{i1}} E_f(t) dt. \quad (14)$$

Here the summation is over the cases that are  $k$  nearest neighbors of focus  $f$ . This statistic will be large when the emission volume of the focus tends to be elevated during times that coincide with the induction periods of its  $k$ -nearest neighbors.

#### **Sensitivity of interaction statistics to specification of the EIP**

At least two instances may arise regarding specification of  $\omega$  and  $\tau$ . The first arises when we are able to model  $\omega$  and  $\tau$  as a function of individual-level characteristics such as genetics, life course, covariates and risk factors. The second arises when we have little knowledge of how  $\omega$  and  $\tau$  may vary from one individual to another. One then may specify  $\omega$  under the simplifying assumption that  $\omega_1 = \omega_2 = \dots = \omega_N$ . The remainder of this paper deals with the second instance, since it is more generally applicable in the absence of the ability to directly observe  $\omega$  and  $\tau$ , and since models of induction period as a function of genetics, risk factors and covariates are typically not available. Given a model of EIP, we follow these steps to assess sensitivity of the interaction statistics.

1. Define the model of EIP and the values of the parameters to explore.

a. Example: For the bladder cancer study we will explore 110 combinations of the induction (1,3,5,7,9,11,13,15,17,19) and latency (5,7,9,11,13,15,17,19,21,23,25) periods.

2. For each parameter set evaluate the distribution of the test statistics under the null hypothesis.

a. Under the null hypothesis of no association between residential history and age at diagnosis allocate the ages at diagnosis with equal probability across the residential histories, calculating the tests for interaction each time. This step is repeated 999 times to generate the distribution of the test statistic under the null hypothesis. For Janus one uses a conditional randomization that keeps the date of diagnosis for the case being considered the same (not randomized). For the Janus statistic, which is a local test, the randomization is conditional in the sense that the date of diagnosis for the case being considered is held constant to be the observed date of diagnosis for that case. The dates of diagnosis for the remaining cases are randomized.

b. Compare the value of the test statistic for the original data to the distribution of the test statistic under the null hypothesis from step 2a. A p-value for a given statistic is calculated for each parameter set.

3. One then inspects the p-values of the global Vesta to identify induction and latency periods that result in significant global interaction. The local statistics may then be used to identify those locations and times contributing the most to the significant global interaction.

#### **The diagnostic process**

A diagnostic process identifies those induction periods and latencies that maximize clustering in exposure traces, while also ameliorating multiple testing (Figure 3). We first use the probability of the global Vesta to assess whether a given latency and induction period is significant (Figure 3, "Global interaction in exposure traces?"). This step is repeated for all sets of induction and latency periods being considered. If none are significant, we advocate for the analysis to cease. While local clustering may be significant [23], as a strategy for ameliorating multiple testing, we only advise searching for those local clusters if the signal is strong enough to also produce a significant global cluster statistic. Those global Vesta statistics (if any) that result in significant global interaction are retained (Figure 3, "At what  $\omega$ ,  $\tau$ ?"), and used to identify the cases, residential locations and times when significant local interaction occurred (Figure 3, "Over whose life course?"). Finally, Janus is applied to identify the locations and times of significant spatial clustering in exposure traces (Figure 3, "When and where do ET cluster spatially?").

#### **The bladder cancer data set**

A population-based bladder cancer case-control study is underway in southeastern Michigan and was used in both simulated and applied studies. Cases diagnosed in the years 2000–2004 and living in Genesee, Huron, Ingham,

Jackson, Lapeer, Livingston, Oakland, Sanilac, Shiawassee, Tuscola, and Washtenaw counties are being recruited from the Michigan State Cancer Registry. Controls from this study are used by us to quantify the probability of being a case given risk factors and covariates. Controls are being frequency matched to cases by age ( $\pm 5$  years), race, and gender, and are being recruited using a random digit dialing procedure from an age-weighted list. At this stage of recruitment, controls are not adequately matched; therefore, age, race, and gender are adjusted for in the analyses. To be eligible for inclusion in the study, participants must have lived in the eleven county study area for at least the past 5 years and have had no prior history of cancer (with the exception of non-melanoma skin cancer). Participants are offered a modest financial incentive and research is approved by the University of Michigan IRB-Health Committee. The data analyzed here are from 374 cases and 490 controls. Refer to [24] for details on geocoding residential histories.

**The simulation study design**

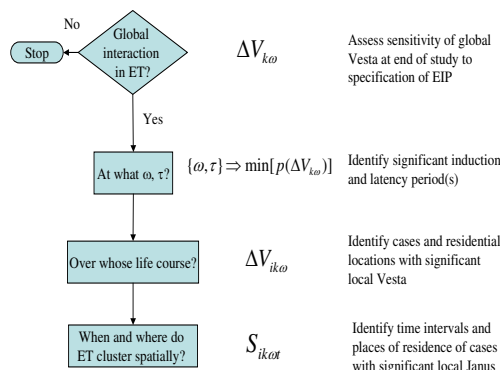
To evaluate type I and type II error we undertook simulations using the residential histories of the cases from the bladder cancer study, but assigned new times of diagnosis based on different scenarios for which the modeled degree of interaction was under experimental control. In each of our experiments we explored sensitivity of the results to pair-wise combinations of induction (1, 3, 5, 7 and 9 years) and latency (5, 7, 9, 11, 13, 15, 17 and 19 years). Three scenarios were analyzed using  $k = 1$  and  $k = 5$  nearest neighbors.

*1) No interaction*

This scenario explored the type I error of the global statistic and the sensitivity of the type I error to specification of induction period and latency. We arbitrarily assigned each case a new date of diagnosis drawn from a uniform distribution between 1990 and 2005, resulting in a dataset without space-time interaction. We then plotted the probability of the global Vesta as a function of the induction and latency periods. This allowed us to evaluate the sensitivity of the global statistic to specification of these parameters when the null hypothesis was true.

*2) Cluster of Size 10*

We modeled a local exposure in early 1985 that resulted in cancers in the exposed group with an induction period of 1 year and a latency of 15 years, resulting in peak years of diagnosis in 1999–2000. We swapped the diagnosis dates for the exposed group with randomly selected members of the remaining cases whose dates of diagnosis were in 1999–2000. This maintained the distribution of dates of diagnosis, and corresponds to an ephemeral exposure of brief duration.



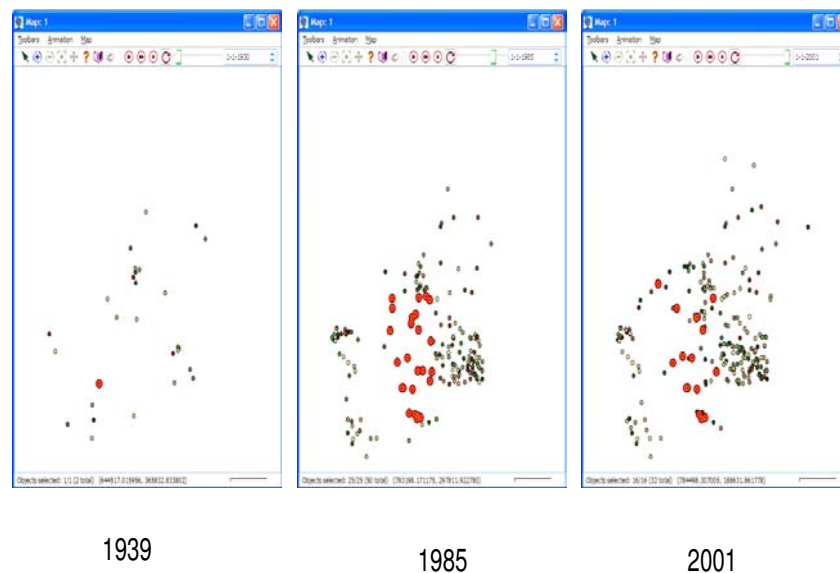
**Figure 3**  
Diagnostic process for exposure traces, see text.

*3) Cluster of size 25*

We modeled a cluster of size 25 occurring in 1985 and incorporating members of cluster size 10 (Figure 4). The induction period (1 year) and latency period (15 years) were maintained.

**Analysis of bladder cancer in Michigan**

Once we had obtained a clearer understanding of the statistical performance and sensitivity of the new methods we applied them to the cases from the bladder cancer study using the original dates of diagnosis. We evaluated  $k = 1$  and  $k = 5$ , but increased the range of the parameters considered for the induction and latency periods. We plotted the probability of the global statistic as a function of the EIP, and for that induction and latency period that resulted in significant global interaction inspected maps of the local statistics to identify clusters of high space-time interaction through time. We then adjusted the tests for known risk factors (smoking) and covariates using the methods described in equation 6. Comparison of the graphs of the probability of the global Vesta as a function of EIP and maps for the tests before and after adjustment allowed us to identify (1) possible contributions of the risk factors and covariates to the induction and latency periods and (2) those local clusters that cannot be explained by smoking and covariates. Clusters that cannot be explained by known factors are of particular interest, as they may be caused by exposures that were not assessed in the case-control study.



**Figure 4**

Evolution of the cluster of size 25. Locations of place of residence of cluster members are shown as red circles in 1939 (left), during the exposure in 1985 (center) and in 2001 (right).

## Results

### Simulation study

#### No Interaction

The plot of the probability of the global Vesta as a function of the parameter values has a minima at 0.107 and a maxima near 1. At an alpha level of 0.05, one would correctly conclude there was no space-time interaction.

We then calculated the values of each of the local statistics through time, and evaluated their significance at each unique arrangement of places of residence. This allowed us to construct graphs of the observed proportion of local statistics that were correctly classified as "not clustered" as a function of the decision criteria for the test. We inspected curves for each parameter set. The correct decision of no interaction is achieved 100% of the time up to a decision level for the test of over 30%. For the scenario considered, the risk of false positives is zero and does not increase until the alpha level of the test is above 0.3.

#### Cluster Size 10

We applied the global Vesta from Equation 8, repeating the analysis for each of the 40 parameter sets. We then plotted its probability as a function of the EIP. A minimum p-value of 0.034 was observed at an EIP of 16 years, corresponding to induction period 1 year and latency of 15 years, the same induction and latency used when modeling the cluster.

We next used the local Vesta to identify those cases experiencing significant interaction over their life course, and the local Janus statistic to find those times when exposure traces clustered. Even though the modeled cluster was ephemeral and small (10 cases), the Vesta and Janus statistics correctly identified its timing, the induction and latency periods used, and found 5 of the cases in the modeled cluster.

#### Cluster Size 25

The sensitivity analysis to specification of EIP found minimum  $p < 0.01$  for the global statistic for an average induction period of 2.7 years and an average latency of 14.7 years, near that of the modeled cluster. The Janus statistic correctly localized the cluster in time, and identified 21 members of the cluster, with 4 false negatives and no false positives. The approach thus appears capable of estimating with acceptable accuracy the latency, induction periods and membership of the simulated clusters.

#### Bladder cancer

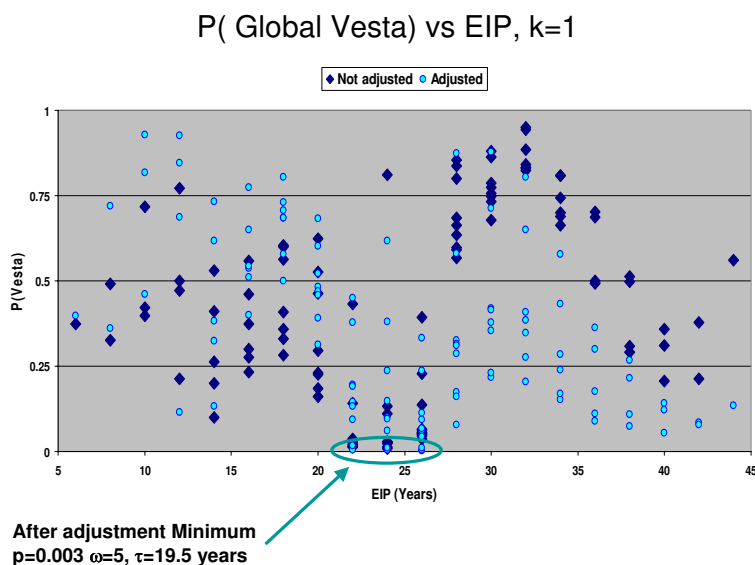
We next analyzed the bladder cancer data to better understand how this new approach might be applied to real data. We analyzed a total of 110 parameter sets using induction periods 1, 3, 5, 7, 9, 11, 13, 15, 17, 19 and latencies 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 and 25 years. This resulted in EIP's from 6 to 44 years. We employed logistic regression and the case and control data to quan-

tify the probability of being a case given the risk factor smoking and the covariates age, gender, education and race (for further description of the logistic model see [22]). We then ran the analyses taking into account these case probabilities, employing the method of Equation 6, and undertook the same analyses without covariate adjustment. We evaluated  $k = 1$  and  $k = 5$  to explore scale dependencies in case clustering. The results using  $k = 5$  were not statistically significant, but were for  $k = 1$ . After adjustment, the smallest probabilities of the global Vesta were for EIP's from 22 to 26 years (Figure 5), with a minima of  $p = 0.003$  occurring at average induction period 5 years, latency 19.5 years. We used these as input to Janus to evaluate local spatial clustering of exposure traces through time. Significant clustering of exposure traces begins in 1975 and continues through 1986 (Figure 6).

**Discussion**

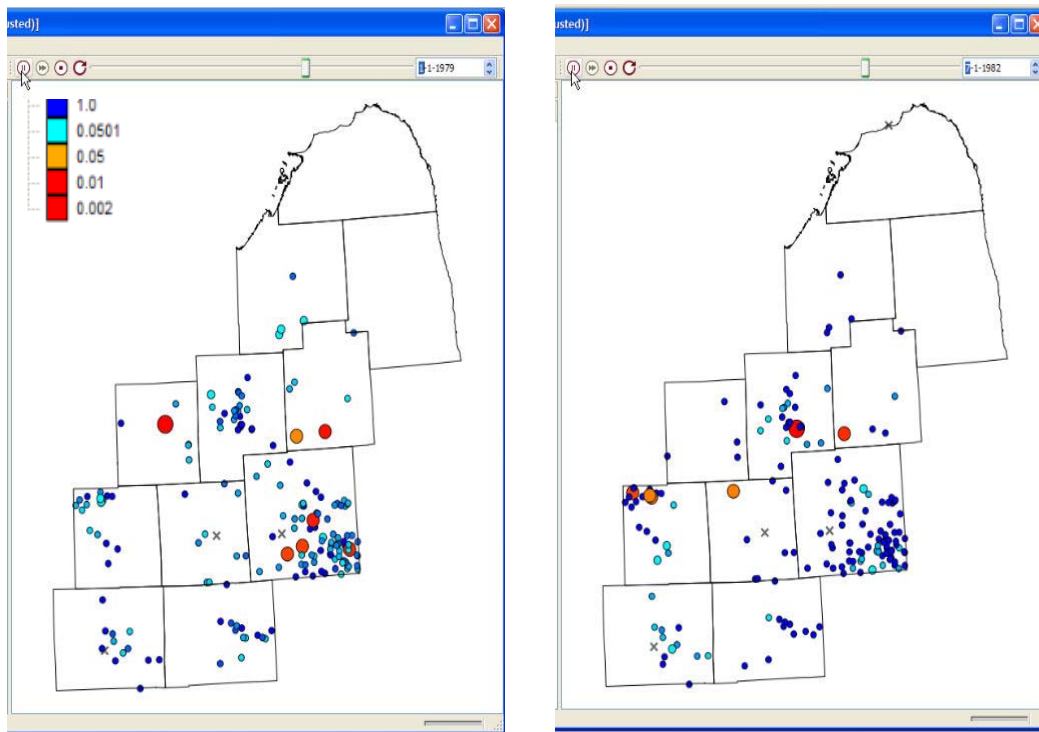
The effects of latency as described in current epidemiological literature are often insufficient to address public health questions, largely because quantitative models of latency are lacking [24]. Langholz et al. [24] developed latency models based on bilinear and exponential decay functions, and fitted these models to case-control data

within a likelihood framework. They defined latency as the function describing how the relative risk associated with a *known exposure* changes through time, and the function may be estimable in occupational studies. As an example, they observed that "... relative risk associated with exposure increases for about 8.5 years and thereafter decreases until it reaches background levels after about 34 years" in a study of lung cancer in a cohort of uranium miners. In contrast, Janus and Vesta evaluate whether the residential histories of cases exhibit interaction during the induction periods – those times when causative exposures plausibly might have occurred – but *we do not necessarily know what those exposures might be*. We thus must use our admittedly inadequate knowledge of cancer latency to define induction periods within which an environmental exposure *might* be causally associated with a given case. This could indicate, for example, those times in a person's life course when exposures (should they occur) are most likely to cause cancer. Several authors have suggested, that when faced with uncertainty, one should explore sensitivity of the latency-based statistic to plausible specifications of the induction period [21,25], and that is the approach used in this paper.



**Figure 5**

Empirical Induction Period sensitivity analysis, bladder cancer study,  $k = 1$ . The probability of the global statistic for space-time interaction is on the y-axis, the x-axis is the EIP in years used when evaluating the global statistic. A minimum of  $p = 0.003$  is reached at an average induction period of 5 years, and a latency of 19.5 years.



**Figure 6**

Local spatial clustering of exposure traces for bladder cancer cases. Shown are the locations of significant clusters for the Janus statistic on 1/1/1979 (left) and 7/1/1982 (right).

The Janus statistic is sensitive to ephemeral spatial clustering of exposure traces, and the simulation studies found that it can pick up the signal from a cluster of brief duration. The Vesta statistics are accumulated over the induction periods, and identify cases who were in close geographic proximity to other cases during their induction periods. The global Vesta thus evaluates interaction in exposure traces at specific induction and latency periods. When interaction is absent the simulations found the global Vesta not significant even when a large number of values of the induction and latency periods are considered. Hence adjustment for multiple testing may not be required to correct the type I error when evaluating a range of empirical induction periods, provided one uses the diagnostic process and first evaluates whether the global statistics are significant before proceeding. Additional simulation studies are needed to evaluate whether this holds over a range of scenarios.

As noted earlier, the simulations we conducted are limited, and it may very well be that false positives will arise under other simulated conditions. Given the simulations we have conducted to date, one possible explanation is

that the methods are more prone to type II error than they are to type I error. This kind of a trade off between type I and type II error is observed for many statistical methods. Further simulation studies are needed to more fully explore the trade offs between type I error, type II error, and statistical power.

Statistical significance of the global Vesta is used to determine (1) whether the analysis should proceed, and (2) what induction and latency periods to employ for the local analyses. The diagnostic framework thus is designed to detect "big signals" that will result in statistical significance of the global Vesta. We do not employ corrections for multiple testing of the local Vesta once significance of the global Vesta has been demonstrated; rather we seek to identify those cases and time periods that contribute the most to a significant global test statistic. The validity of this approach is supported by simulation, in which clusters of size 25 and even of size 10 were localized with small type I error, and returned appropriate induction and latency periods. Janus found 5 members of the cluster of size 10 and 21 members of the cluster of size 25, with cases that were missed occurring on the cluster edge. This

seems to be reasonable performance given the small cluster size and the ephemeral nature of the modeled clusters.

When considering multiple testing, Fuchs and Kenett [23] argued, in the aspatial case, that a test of the most extreme local statistic (accounting for multiple testing) can be more powerful at finding clusters than the use of the corresponding global test. This likely may be true for spatial tests as well, in which case significant local clusters might be identified even when the global statistic is not significant.

Several caveats apply to the simulation design. We constructed the simulations to be simple, and yet to pose a fairly stringent "first test" of the new methods by modeling clusters of short duration and size. We decided to swap dates of diagnosis when constructing the clusters, making interaction and clustering of exposure traces the only aspect of the dataset that would change across simulations – the frequency distribution of dates of diagnosis was constant. We used a cluster of size 10 and 1 year duration as the smallest, and were pleasantly surprised to find the methods indeed were sensitive enough to find that cluster. Nonetheless, additional simulations are needed to address the impacts of uncertainty in the residential histories, multiple clusters, and of heterogeneity in individual induction and latency periods.

In order to generate bias in interaction of the exposure traces one would need to preferentially sample a subset of the population with similar dates of diagnosis that were in geographic proximity to one another during their induction periods. This might occur for rural populations characterized by little residential mobility. At first blush a second potential source of bias might be differential mobility in different parts of the study area. Localities with greater residential mobility might have larger variability in the temporal overlap of exposure traces, since individuals on average do not stay as long in any given place of residence. The randomization procedure holds the residential histories as a given, and permutes dates of diagnosis across the cases. Differential residential mobility should therefore be accounted for under the null hypothesis. Finally, changes in diagnostic procedures such that risk of diagnosis increases at different times in different parts of the study area are a potential source of bias, since this would lead to an apparent overlap in exposure traces. This would definitely create clustering at time of diagnosis, but we'd expect the cluster to become diffuse by time of the induction period due to residential mobility, unless the induction period is close to time of diagnosis.

At the time this article was written the bladder cancer study was in progress and cases were still being enrolled.

A portion of the thumb of Michigan – those counties in the North of the study area – have yet to be visited by the field teams for the latest round of sampling. These comprise a primarily rural population with recent dates of diagnosis, a potential source of sampling bias (i.e., differential sampling across the study area) that could result in spurious findings of significant interaction. We thus must wait before attaching further interpretation to clusters of exposure traces found under the Janus and local Vesta statistics.

What is the reason for this differential sampling? For the bladder cancer study differential sampling arose because of the timeline chosen for household visits to residences of the cases and controls. These visits included survey instruments, water sampling to assess arsenic concentrations in the water supply, and biological sampling such as toenail clippings and buccal samples to assess recent arsenic exposure and genetic factors. Many of these sample instruments and assays were tangential to the topic of the current paper, and are discussed in detail in other peer-reviewed publications. Differential sampling at the time of this writing arose because sampling is systematic geographically in order to reduce expense – the sampling team goes into an area (say the southern part of the study area) and visits those residences, at a later date visits residences in another area, and so on. Hence while the overall sample is representative, the manner in which the data are collected is geographically and temporally sequential. Thus when we analyze data before data collection is complete our sample up to that point in time necessarily is differential. This of course will not be an issue when we conduct analyses after data collection is finished.

If these clusters persist once data collection is complete, we will need to investigate environmental agents hypothesized to cause bladder cancer that produce an induction period of five years, followed by a latency period of nearly twenty years. In addition, the agent or agents responsible only resulted in clusters using one nearest neighbor, not the nearest five neighbors, suggesting tight geographic areas of high exposure. One might conjecture that a possible cause of this space-time clustering pattern is pollution from several local industries in the region [12], or a more disperse contaminant that appears in localized hotspots, such as arsenic in private well water which is found in elevated concentrations in southeastern Michigan [26]. Examination of these hypotheses will involve thorough exposure assessment; however, the space-time clustering approach introduced here can help bring these possible exposures to light. These analyses will be repeated once data collection is complete.

The strength of the Janus and Vesta statistics lies in their ability to help identify induction and latency periods, an

area of research generally underserved in cancer epidemiology. Most efforts aimed at understanding the temporal relationship between exposure and cancer have focused on improving the temporal resolution of exposure assessments [26-28]. In this paper, we take advantage of disease and residential history datasets for gaining insights about the temporal dynamics of the exposure-disease relationship. We developed statistics for quantifying space-time interaction in exposure traces, while allowing the user to explore a range of induction and latency periods. If clustering is identified after following this approach, this calls for investigation into temporally characterized exposures potentially responsible for the clustering.

These new methods raise the possibility of routine surveillance using cancer registry data in a manner that accounts for individual mobility, identifies plausible values of the induction and latency periods, and that identifies geographic locations and times associated with increases in cancer risk *above and beyond* that expected given known covariates and risk factors in geographically mobile populations.

### Authors' contributions

GJ created the Vesta and Janus statistics, undertook the simulation experiments and bladder cancer analyses, and drafted the ms. JM identified the need for interaction statistics for mobile populations, undertook the regression analyses for specifying the case probabilities, and coordinated preparation of the bladder cancer data. AK programmed the methods in the Space-Time Intelligence System. All authors read and approved the final manuscript.

### Acknowledgements

This research was funded by grants R43CA117171, R43 CA112743 and R01CA096002 from the National Cancer Institute. The views expressed in this publication are those of the researchers and do not necessarily represent that of the NCI. The authors thank two anonymous reviewers for their comments and suggestions.

### References

- Bell BSHR, Pickle LW, Wartenberg D: **Current practices in spatial analysis of cancer data: mapping health statistics to inform policymakers and the public.** *International Journal of Health Geographics* 2006, **5(49):**.
- Waller LA, Jacquez GM: **Disease models implicit in statistical tests of disease clustering.** *Epidemiology* 1995, **6(6):**584-590.
- Besag J, Newell J: **The detection of clusters in rare diseases.** *Journal of the Royal Statistical Society* 1991, **Series A(154):**143-155.
- Cuzick J, Edwards R: **Spatial clustering for inhomogeneous populations.** *Journal of the Royal Statistical Society* 1990, **Series B(52):**73-104.
- Kulldorff M, Nagarwalla N: **Spatial disease clusters: detection and inference.** *Stat Med* 1995, **14(8):**799-810.
- Kulldorff M, Song C, Gregorio D, Samociuk H, DeChello L: **Cancer map patterns: are they random or not?** *Am J Prev Med* 2006, **30(2 Suppl):**S37-49.
- Tango T, Takahashi K: **A flexibly shaped spatial scan statistic for detecting clusters.** *Int J Health Geogr* 2005, **4:**11.
- Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC: **Monitoring for clusters of disease: application to leukemia incidence in upstate New York.** *Am J Epidemiol* 1990, **132(1 Suppl):**S136-143.
- Waller LA, Turnbull BW: **The effects of scale on tests for disease clustering.** *Stat Med* 1993, **12(19-20):**1869-1884.
- Waller LA, Turnbull BW, Gustafsson G, Hjalmars U, Andersson B: **Detection and assessment of clusters of disease: an application to nuclear power plant facilities and childhood leukemia in Sweden.** *Stat Med* 1995, **14(1):**3-16.
- Jacquez GM: **Current practices in the spatial analysis of cancer: flies in the ointment.** *Int J Health Geogr* 2004, **3(1):**22.
- Jacquez GM, Kaufmann A, Meliker J, Goovaerts P, AvRuskin G, Nriagu J: **Global, local and focused geographic clustering for case-control data with residential histories.** *Environ Health* 2005, **4(1):**4.
- Sabel CE, Boyle PJ, Loytonen M, Gatrell AC, Jokelainen M, Flowerdew R, Maasilta P: **Spatial clustering of amyotrophic lateral sclerosis in Finland at place of birth and place of death.** *Am J Epidemiol* 2003, **157(10):**898-905.
- Mantel N: **The detection of disease clustering and a generalized regression approach.** *Cancer Res* 1967, **27(2):**209-220.
- Knox G: **The detection of space-time interactions.** *Applied Statistics* 1964, **13:**25-29.
- Diggle P, Chetwynd A, Haggkvist R, Morris S: **Second-order analysis of space-time clustering.** *Statistical methods in medical research* 1995, **4:**124-136.
- Jacquez GM: **A k nearest neighbour test for space-time interaction.** *Stat Med* 1996, **15(17-18):**1935-1949.
- Kulldorff M, Hjalmars U: **The Knox method and other tests for space-time interaction.** *Biometrics* 1999, **55(2):**544-552.
- Aldstadt J: **An Incremental Knox Test for the Determination of the Interval between Successive Cases of an Infectious Disease.** *Stochastic Environmental Research & Risk Assessment* 2007, **21:**487-500.
- Jacquez GM, Meliker JR: **Geographic Clustering for Mobile Populations.** In *Handbook of Spatial Analysis* Edited by: Fotheringham S, Rogerson P. Sage Publications; 2007.
- Rothman KJ: **Induction and latent periods.** *Am J Epidemiol* 1981, **114(2):**253-259.
- Jacquez GM, Meliker JR, AvRuskin GA, Goovaerts P, Kaufmann A, Wilson M, Nriagu J: **Case-control geographic clustering for residential histories accounting for risk factors and covariates.** *International Journal of Health Geographics* 2006, **5(32):**.
- Fuchs C, Kenett : **A test for detecting outlying cells in the multinomial distribution and two-way contingency tables.** *Journal of the American Statistical Association* 1980, **75:**395-398.
- Langholz B, Thomas D, Xiang A, Stram D: **Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado Plateau uranium miners cohort.** *Am J Ind Med* 1999, **35(3):**246-256.
- Meliker JR, Jacquez GM: **Space-time clustering of case-control data with residential histories: Insights into empirical induction periods, age-specific susceptibility, and calendar year-specific effects.** *Stochastic Environmental Research & Risk Assessment* 2007, **21:**625-634.
- Meliker JR, Slotnick MJ, AvRuskin GA, Kaufmann A, Fedewa SA, Goovaerts P, Jacquez GM, Nriagu J: **Individual lifetime exposure to inorganic arsenic using a Space-Time Information System.** *International Archives of Occupational and Environmental Health* 2007, **80:**184-197.
- Agalliu I, Eisen EA, Kriebel D, Quinn MM, Wegman D: **A biological approach to characterizing exposure to metalworking fluids and risk of prostate cancer (United States).** *Cancer Causes and Control* 2005, **16:**323-331.
- Palmer VW Jr, Hatch EE, Troisi R, Titus-Ernstoff L, Strohsnitter WW, Kaufman R, Herbst AL, Noller KL, Hyer M, Hoover RN: **Prenatal diethylstilbestrol exposure and risk of breast cancer.** *Cancer Epidemiol Biomarkers Prev* 2006, **15:**1509-1514.