5-2016

A Bayesian Method for Cluster Detection with Application to Five Cancer Sites in Puget Sound

Albert Y. Kim
Reed College, akim04@smith.edu

Jon Wakefield
University of Washington

Published in final edited form as:
Epidemiology. 2016 May ; 27(3): 347–355. doi:10.1097/EDE.0000000000000450.

A Bayesian method for cluster detection with application to five cancer sites in Puget Sound

Albert Kim[1] and **Jon Wakefield**[2]

[1]Mathematics Department, Reed College, Portland, OR 97202

[2]Departments of Statistics and Biostatistics, University of Washington, WA 98112

Abstract

Cluster detection is an important public health endeavor and in this paper we describe and apply a recently developed Bayesian method. Commonly-used approaches are based on so-called scan statistics and suffer from a number of difficulties including how to choose a level of significance and how to deal with the possibility of multiple clusters. The basis of our model is to partition the study region into a set of areas which are either "null" or "non-null", the latter corresponding to clusters (excess risk) or anti-clusters (reduced risk). We demonstrate the Bayesian method and compare with a popular existing approach, using data on breast, brain, lung, prostate and colorectal cancer, in the Puget Sound region of Washington St ate. We address the important issues of sensitivity to the priors, and the incorporation of covariates. The approach is implemented within the freely-available `R` package `SpatialEpi`.

Introduction

Cluster detection has a long and controversial history in epidemiology. Unfortunately, only very rarely have etiological insights been made as a result of investigations into clustering and clusters[1] which has led Rothman[2] to call this endeavor into question. Neutra,[3] in a response to Rothman, agreed with a number of his conclusions but believed that investigation of putative clusters was a necessary part of the public health response; cluster detection can aid in identification of areas in need of resources or interventions such as public awareness campaigns and screening.

In this paper, we describe a new approach to cluster detection, based on a Bayesian model, and describe the use of the method for cluster detection for five cancers in the Puget Sound region of Washington State. First, we briefly review a number of the previous approaches to cluster detection that have been proposed. The most popular approaches are based on *scan statistics*, in which a circular window is passed over the study region and the significance of the observed number of cases in the window is determined. Different proposals base the circle size on distance,[4] the number of cases[5] and on the population.[6] The latter method has been extensively refined and forms the basis of the method implemented in the `SatScan` software (http://www.satscan.org/.[7]) The model behind this approach assumes circular clusters are centered on each of the n area centroids. Circles with varying radii, up to a maximum that gives a circle with no more than a certain proportion of the total population (20% is a common choice) are considered. We refer to the circles as (single) *zones*. Hence, any one area will typically fall within a large number of potential zones. Hypothesis testing is used to determine the significance of clusters by comparing the observed and expected numbers of disease under the null hypothesis. In the version relevant to the application considered here, a Poisson likelihood is assumed and a likelihood ratio statistic is calculated for each zone, with the null corresponding to no clusters. This approach clearly leads to a large number of dependent tests, and the multiplicity problem is circumvented by evaluating the significance of only the *maximum* of the likelihood ratio statistics over all circles, using a Monte Carlo p-value. The Monte Carlo p-value is computed by comparing the observed test statistic to a simulated null distribution. Each instance of the simulated null distribution is constructed by randomly assigning cases to locations under the null hypothesis of no clusters and computing the test statistic.

There are a number of drawbacks to the SatScan method. First, as with all frequentist testing procedures, a fundamental problem is how to interpret the resultant p-value and in particular choose a threshold below which “significance” should be declared. These difficulties are well-documented.[8–10] A specific problem with p-values is that their interpretation critically depends on the power of the study. In the context of cluster detection, $\alpha = 0.05$ was used in both Kulldorff et al.[11] and Jemal et al.[12] In the former, breast cancer was examined over 245 counties of the North East of the US with 58,943 deaths and a population of 29,535,216. In the latter, prostate cancer was examined in males in the United States in 1970–1989 with 71,692 black deaths and 382,204 white deaths amongst the total US male population. Consequently, the power is very different in the two studies but there is no explicit consideration of power in the calculation and interpretation of p-values; stated more bluntly, there is no balancing of type I and type II errors. Intuitively, the p-value significance threshold should decrease as the sample size increases since the power is increasing, i.e. the type II error is decreasing, and so the type I error should be decreasing also. A Bayesian approach to inference acknowledges that the type I and type II errors should go to zero and provides the machinery to avoid these problems. The standard Bayesian way to evaluate the evidence in the data for particular null and alternative hypotheses H_0 and H_A is the Bayes factor, given by $p(\text{data}|H_0)/p(\text{data}|H_A)$. The numerator and denominator of the Bayes factor represent, respectively, the probability density of the data under the null and alternative hypotheses, and these calculations depend on the assumed model. The Bayes factor accounts for power through its denominator, and type I and type II errors are naturally balanced in the numerator and denominator.[9;10;13]

A second difficulty with the SatScan method is how to deal with the possibility of multiple clusters. The original version of the Kulldorff method, simply compared the p-values of the second, third, etc. most significant zones using the reference distribution for the maximum (most likely) cluster, after discarding those with overlap with the first cluster. This approach is therefore not using the correct reference distribution. The most recent multiple cluster version of the Kulldorff method[14] removes the significant zone, and then repeats the procedure now using a new reference distribution, until no more significant zones are found. A deficiency of this approach is that the p-values are not directly comparable since they are based on different sample sizes and hence have different power. Also, since the procedure removes significant clusters from the study region at each iteration, the significance of secondary clusters must be interpreted as conditional on all previously identified significant clusters and not as standalone measures as is desired. Furthermore, one should also consider the multiple testing aspect of the multiple comparisons that are being made but it would be very difficult to determine the appropriate error rate of the overall sequential procedure just described.[14]

We briefly describe previous approaches to Bayesian cluster detection. The usual *smoothing* of rates or relative risks, as implemented in common disease mapping models,[15] is not a good idea, as clusters may be attenuated due to the shrinkage of these models.[16] We have previously described our model for cluster detection[17] and illustrated its use with the classic upstate New York leukemia dataset.[18]

In this paper we make some modest extensions to the methodology and demonstrate the approach in a more extensive application. Specifically, we apply the models to data on female breast, brain, lung, male prostate, and colorectal cancer, collected in the Puget Sound region of Washington State between 1996–2005. We choose breast, lung, prostate and colorectal because they are relatively common, while brain is relatively rare and so was picked to give an example of the methodology for a rarer cancer. Geographical analyses have also revealed excess brain cancer mortality in the northwest of the United States,[19;20] a region that contains our study region. The results for breast cancer require careful interpretation and so we detail these in the main paper, while for brain cancer interpretation is more straightforward, and we include this in the paper as a contrast. The results for the remaining three cancer sites are presented in the Supplementary Materials. Our approach is most similar to that described by Gangnon and Clayton[21–23] though with differences described elsewhere.[17] We have implemented our method in the `R` computing environment within the `SpatialEpi` package.

Brain and Breast Cancer in Western Washington

Study Details

The Cancer Surveillance System (CSS) collects population-based data on cancer incidence and survival in $n = 887$ census tracts in 13 counties in western Washington State: Clallam, Grays Harbor, Island, Jefferson, King, Kitsap, Mason, Pierce, San Juan, Skagit, Snohomish, Thurston, and Whatcom counties. The CSS is a project of the Program in Epidemiology, Division of Public Health Sciences, at the Fred Hutchinson Cancer Research Center (FHCRC) and is part of the Surveillance, Epidemiology, and End Results (SEER)[24] program of the National Cancer Institute which monitors cancer incidence and survival in approximately 26% of the US population. The CSS also provides data to the Washington State Cancer Registry, which monitors cancer incidence in the entire state of Washington.[25] In Figure 1 we present a map of the study region with the Seattle and Tacoma metropolitan regions highlighted with solid and dashed lines respectively and the city of Mount Vernon marked with a dot. All subsequent maps will include focused maps of both Seattle and Tacoma. Furthermore all values in legends are grouped by the Jenks natural breaks classification method which maximizes the variance between the group means while minimizing the within-group variance of values.[26]

One must bear in mind that two of the chief reasons for carrying out a cluster detection endeavor are to discover previously unknown exposures or risk factors that are linked with clusters, or to detect areas with high risks (perhaps due to inaccessibility to health care) in order to prioritize public health resources and interventions (encouragement of screening, for example). If we are interested in the former, then adjustment for as many known risk factors as possible is merited. If we are interested in the latter, then we would not want to adjust for area-level variables such as income, since this will mask important differences. In the study we present here, we are interested in detecting areas with high residual (i.e. after adjustment for risk factors) risk and we report results with and without adjustment for an area-level measure of income.

As notation we will use y_{ij} to represent the disease count in each area $i = 1, \ldots, n$ and confounder stratum $j = 1,\ldots, J$, with N_{ij} the corresponding population size. For each cancer site, counts of disease incidence y_{ij} were obtained for each census tract $i = 1,\ldots, 887$ stratified by $j = 1,\ldots, 180$ strata: age (18 age bands: 0–4, 5–9, 10–14, …, 75–79, 80–84, 85+), race (5 categories: White, Black, Asian/Pacific Islander, American Indian and other including those of two or more races) and gender for brain, lung, and colorectal cancers, while we only consider women for breast cancer and men for prostate cancer. These disease counts were combined across years and paired with corresponding population counts N_{ij} obtained from the 2000 Decennial Census. For each of the five cancer sites, we calculate two sets of expected numbers in area i, denoted E_i, using internal standardization, at the level of the census tract. In one set we adjust for age, race, and gender (when required) and in the other set we additionally adjust for income by quintile. We consider 1999 per capita income as reported by the American Community Survey at the census tract level only, since age, race, and gender stratified income data are not available. Our analysis will consider income through the quintiles defined by cutoffs $18,478, $21,661, $24,825, and $30,161. The income adjusted reference rates of disease for census tract i are based on disease and population counts from census tracts in the same income bracket. The Standardized Morbidity Ratio (SMR) in census tract i is calculated as $\text{SMR}_i = y_i/E_i$. The SMR is an area-based summary and an estimate of the relative risk in an area, when compared to the reference rates that were used to construct the expected numbers.

Table 1 displays summaries of the aggregated disease counts y_i for all 5 cancers, along with expected numbers E_i and standardized morbidity ratios SMR_i, across areas $i = 1,\ldots, n$, unadjusted and adjusted for income. Breast is the most common cancer, followed by prostate, lung, colorectal, and brain. We now present detailed analyses for brain and breast cancer, with lung, colorectal and prostate discussed in the Supplementary Materials.

Initial Analysis for Brain Cancer

Before a cluster detection exercise is carried out it is important to have knowledge of known risk factors, but the cause of many brain cancer cases is unknown (http://www.cancer.gov/cancertopics/types/breast). Previously, there was evidence of excess brain cancer mortality in the northwest United States, as revealed in a cancer mortality atlas,[19] and this was followed up with a formal study of brain cancer mortality across the United States from 1986–1995 using, amongst other statistical techniques, SatScan.[20] This study reported evidence of a cluster in Washington State.

For our study the expected numbers for brain cancer incidence are displayed in Figure 2, with the left and right columns being unadjusted and adjusted for income, respectively. The corresponding SMRs are displayed in Figure 3. Compared to the other cancers the expected numbers are relatively small, but the SMRs show large variability, which may of course be attributed to sampling variability due to the relatively small expected numbers for many census tracts. A plot of (log) SMRs versus income quintile (both with and without adjustment for income), provided in the Supplementary Materials, shows very little evidence of a census tract level association between relative risk and income.

We present results of the multiple cluster Kulldorff method[14] using both income unadjusted and adjusted expected counts. We set the highest proportion of the population a zone can contain at 15%, the α-cutoff to declare significance at $\alpha = 0.05$ and simulate 9,999 Monte Carlo realizations under the null hypothesis of no clusters. The results are presented in Figure 4. For both income unadjusted and adjusted expected counts, there are no clusters that are significant at the 5% level.

Initial Analysis for Breast Cancer

We first briefly summarize the risk factors for breast cancer, leaning heavily on http://www.cancer.gov/cancertopics/types/breast. The lifetime risk for developing breast cancer for women in the United States is approximately 1 in 8. Risk factors for breast cancer include genetic alterations, dense breasts, exposure to estrogen (early menstruation, late menopause, no pregnancy or late pregnancy), family history of breast cancer, alcohol, race (breast cancer is diagnosed more in white women) and obesity. Protective factors for breast cancer include less exposure to estrogen (for example through early pregnancy, breast-feeding, increased number of births and increased duration of breast feeding) and exercise.

The geographical distribution of breast cancer risk has received a reasonable amount of interest with Kulldorff et al.[11] providing an early example of the application of a scan statistic to breast cancer, with numerous other studies following. For example, scan statistics have been used to detect clusters for breast cancer in Massachusetts,[27] Texas,[28] Connecticut,[29] and across all counties in the United States[30]. An interpretation of geographical variation across the United States has also been carried out,[31] with a major conclusion being that at large scales at least, there are differences in presentation (i.e. stage) by region and by race. A large number of geographical analyses have examined the association between clusters and environmental pollution sources (see for example Jacquez and Greiling[32] and references there-in). Interest in cluster detection for breast cancer extends beyond the United States, for example, breast cancer has been examined geographically in Japan, using the Kulldorff scan statistic.[33]

The expected numbers for breast cancer are displayed in Figure 5 with the SMRs displayed in Figure 6. The expected numbers for breast cancer are far greater than for brain and while the SMRs for breast cancer have roughly the same range as with brain, the former are more accurately estimated given the larger expected counts. In the Supplementary Materials we plot the (log) SMRs (without adjustment for income) versus income quintiles, and we see that there is a negative association, with areas in higher income quintiles having higher relative risks. This is not unexpected given the above risk factors for breast cancer since we would expect women living in areas with low average income to have more and earlier pregnancies and a greater duration of breast feeding.

Figure 7 gives the results of the scan statistic both without and with adjustment for income. Without income there is one very large cluster in Seattle that has a significance level < 0.0001, with three additional clusters, one to the west of Seattle, one in Tacoma and a single census tract near Mt Vernon in the north of the study region. With adjustment for income there are three areas, two in Tacoma with estimated relative risks of 5.671 and 2.320, and a third consisting of the single census tract near Mt Vernon with an estimated relative risk 1.146.

A Bayesian Cluster Detection Model

As with the scan statistic method, suppose the study region is partitioned into $i = 1,\ldots, n$ *areas*, which are typically administrative subdivisions of a region such as census tracts, zip codes, or counties. Let θ_i be the relative risk in area i, where relative is with respect to the reference rates for disease that were used to construct the expected numbers. The Supplementary Materials contain a mathematical description of our model; here we give an explanation in heuristic terms.

Configurations as Data Models

Following Kulldorff[6] we define what we refer to as *single zones*: contiguous collections of areas that form "jagged circles." We create the list of single zones by sequentially aggregating neighboring areas, by taking each area in turn, and continually adding the areas whose centroids are closest to the area center. This procedure is continued until the zone's population reaches a pre-specified maximum allowable proportion of the total study region's population, typically under 50%. Our cluster detection model treats each of the N_1 resulting single zones as a potential cluster (region of high residual risk) or anti-cluster (region of low residual risk).

First, suppose that there exists no more than one cluster/anti-cluster in the study region. We assume the data can be explained by $N_1 + 1$ possible configurations, where each configuration can be viewed as a model for $y_1,\ldots, y_n$. The first is a null configuration of no clusters/anti-clusters which assumes that all areas in the study region have "null" relative risks that are close to 1. The remaining N_1 configurations assume that one and only one single zone is a cluster/anti-cluster where within a single zone all areas share a common "non-null" relative risk, i.e., elevated or lowered risk, while all areas outside the single zone have null relative risks. In null areas we do not force the relative risks to be exactly 1, but rather assume that the relative risks arise from a "narrow" distribution that is concentrated close to 1, reflecting the fact that even with true null risk unmeasured confounders and data anomalies can still yield some variability around 1. In contrast, for non-null areas within a single zone, we assume they share a common relative risk that arises from a "wide" distribution that is more diffuse, though still centered around 1. An illustrative example of null/narrow and non-null/wide relative risk distributions using two gamma distributions is given in Figure 8 where the respective 95% intervals are marked with dashed lines. We discuss the specification of the gamma distribution parameters in the next section.

We now describe how the framework is extended to allow for the possibility of more than one cluster/anti-cluster. Let N_j for $j = 2,\ldots, J$ be the number of combinations of j single zones that do not overlap, up to a pre-specified maximum of non-null regions J (hence, the maximum number of clusters/anti-clusters is J). For our examples in this paper we specify a maximum of $J = 7$ clusters/anti-clusters in the study region. Given that we only consider non-overlapping single zones, we assume that the relative risks in non-overlapping single zones are independent and arise from the wide distribution and, as with the single cluster/anti-cluster case, all areas not included in a single zone have null relative risks. These configurations are the (discrete) unknown parameter of our cluster detection model; we denote this parameter c. The null configuration is denoted $c = \{0, 1\}$ and $c = \{j, k\}$ denotes a non-null configuration k consisting of j single zones for $k = 1,\ldots, N_j$. An example of a configuration consisting of $j = 2$ single zones can be found in the Supplemental Materials. The posterior probability $\Pr(c|y_1,\ldots, y_n)$ for all configurations c is a summary of interest of our model and is computed using Bayes theorem.

There are $1 + N_1 + N_2 + \ldots + N_J$ possible configurations, that is models for the data, to consider. This value is very large in typical applications. For example, for the SEER data there are $n = 887$ census tracts and $N_1 = 117, 006$ single zones. The number of multiple single zone configurations then grows quickly and we can only estimate, rather than enumerate, these quantities. For example $\hat{N}_2=5.3\times10^9$, $\hat{N}_3=1.2\times10^{14}$, and $\hat{N}_4=1.6\times10^{18}$. Hence, it is not computationally feasible to enumerate all possible configurations and compute their exact posterior probabilities. Therefore, as an alternative we search through the space of all possible configurations using a Markov chain Monte Carlo (MCMC) algorithm to approximate all posterior probabilities. Details of the computation are in the Supplementary Materials, and the method is implemented within the `SpatialEpi` package within the `R` computing environment.

Likelihood and Prior

A typical choice of model for rare diseases is to assume, for a generic region with count y, $y|\theta \sim \text{Poisson}(E\theta)$ where θ is the relative risk associated with the region and E is the expected number of disease.[34] By assuming a conjugate Gamma(a, b) prior on θ with shape and rate parameters (a, b) the marginal likelihood $\Pr(y)$ is Negative Binomial $\left(a, \frac{b}{E+b}\right)$. As previously described we have two specifications of the Gamma parameters (a, b) corresponding to the null and non-null distribution of relative risks described earlier: a narrow specification $(a_{\text{n}}, b_{\text{n}})$ and a wide specification $(a_{\text{w}}, b_{\text{w}})$. These values are chosen by specifying two θ points with a pair of characteristics. In particular, we set the mode of the distribution to be $\theta = 1$, a value reflecting no increased nor decreased risk, and we specify the 95^{th} percentile of the distribution (see the Supplementary Materials for further discussion). For example, specifying 95^{th} percentiles of 1.03 and 4, we have Gamma parameters of $(a_{\text{n}}, b_{\text{n}}) = (2976.30, 2977.30)$ and $(a_{\text{w}}, b_{\text{w}}) = (2.31, 1.31)$ respectively, and these are the choices shown in Figure 8. Furthermore the points (θ_L, θ_H) at which the two gamma distributions intersect are used as thresholds to declare an area's relative risk as being elevated/reduced since outside the interval defined by these two points the wide specification has higher density. For the above choices of gamma priors the crossover points are (θ_L, θ_H) = (0.949, 1.052). For the null configuration c = {0, 1} all disease counts y_i are independent with relative risks θ_i from the narrow specification for i = 1,…, n. For all non-null configurations c = {j, k}, since the j single zones are independent, the likelihood for the data is the product of the likelihoods associated with the j single zones and the likelihoods of all remaining null areas.

We require a prior on the configurations c = {j, k}, with the number of clusters/anti-clusters being j, and k indexing the configurations for that j. We assign a mass of π_0 = Pr(c = {0, 1}) for the prior on the null, with a typical figure being 0.95 or 0.99, since a priori we expect the chance of clusters to be small. The remaining mass of 1 − π_0 needs to be spread over the remaining configurations c = {j, k} for j = 1,…, J, k = 1,…, N_j. Within each j, i.e. for each number of clusters/anti-clusters, we take each configuration to be equally likely, i.e.

$\Pr\left(c=\{j,\ k\}\left|\tau=j\right.\right)=\frac{1}{N_j}$, where τ is the number of clusters/anti-clusters. It only remains to distribute 1 − π_0 over j = 1,…, J. Recall that each set of j clusters within a particular configuration must be non-overlapping. For a particular j, let q_j be the probability that a randomly selected set of j clusters are non-overlapping. For example, for j = 2, q_2 is the probability that two of the randomly selected N_2 possible single zones are non-overlapping. One may then take

$$\Pr\left(\tau=j\right)=\left(1-\pi_0\right)\times\frac{q_j}{\sum_{l=1}^{J}q_l},$$

so that the marginal prior on the number of clusters/anti-clusters is implied by our choice of the prior on the null, and the set of decreasing probabilities on increasing numbers of clusters/anti-clusters. This produces a set of prior probabilities that decrease monotonically from J = 1, in a relatively natural fashion. As an illustration, with the Puget Sound study geography, the probabilities for j = 0,…, J we obtain from this prescription are 0.950, 0.034, 0.013, 0.003, 2.941 × 10^{-4}, 1.993 × 10^{-5}, 1.027 × 10^{-6}, and 3.029 × 10^{-8}. There are many ways we could specify the set of J + 1 probabilities, and we encourage examining the sensitivity of the results to different choices (see later for examples). For example, we may also spread the probability of a non-null configuration 1 − π_0 equally over j = 1,…, J.

Posterior Probabilities

Each configuration c = {j, k} for j ≥ 1 is made up of j single, non-overlapping zones. The posterior probability for the null configuration is simply

$$\Pr(c=\{0,1\}|y_1,\dots,\ y_n)\propto\pi_0\times\prod_{i=1}^{n}\Pr(y_i|\text{null}).$$

For $j \geq 1$:

$$\Pr(c=\{j,k\}|y_1,\ldots,y_n) \propto (1-\pi_0) \times \prod_{i \in c}\Pr(y_i|\text{non- null}) \times \prod_{i \notin c}\Pr(y_i|\text{null})$$

where $i \in c$ denotes all the areas included in any of j single zones that form configuration $c = \{j, k\}$. In other words, all areas included in any of the single zones follow the non-null model where areas within a common single zone share a common relative risk from the wide prior and all remaining areas are independent with relative risks from the narrow prior. See Supplementary Materials for more details on this derivation.

It is difficult to make statements on the occurrence of clusters/anti-clusters based on the posterior probabilities of *individual* configurations given the large number of configurations to consider, many of which are only minor variations of each other. As an alternative measure, we may calculate the posterior probability of cluster membership for each area i. For a non-null configuration $c = \{j, k\}$ with area i included in one of the j single zones, we evaluate the posterior probability of that single zone's relative risk being "elevated", where elevated indicates higher than the high crossover point θ_H. Analogously we can define the posterior probability of anti-cluster membership for area i using the lower crossover point θ_L. Furthermore, before seeing the data we can evaluate the prior probability of cluster membership for each area i using the prior distribution of the relative risk. Another summary measure of interest is the posterior probability of the number of clusters/anti-clusters in the data. For $\tau = j$ clusters/anti-clusters we have

$$\Pr(\tau=j|y_1,\ldots,\mathrm{y}_n)=\sum_{k=1}^{N_\tau}\Pr(c=\{\tau,k\}|y_1,\ldots,\ y_n)$$

which sums the posterior probabilities of all configurations with τ single zones. Of particular interest will be the $\tau = 0$ case corresponding to no clusters/anti-clusters.

Results

Bayesian Analysis for Brain Cancer

We perform the Bayesian analyses with the maximum proportion of the population a single zone can contain being 15%, wide $(a_\text{w}, b_\text{w}) = (2.31, 1.31)$ and narrow $(a_\text{n}, b_\text{n}) = (2976.30, 2977.30)$ specifications of the Gamma parameters, as shown in Figure 8, a prior probability $\pi_0 = 0.95$ of no clusters/anti-clusters, and a maximum number of clusters/anti-clusters of $J = 7$.

In Figure 9 we present the results for brain cancer using the income unadjusted expected counts in the left column and income adjusted expected counts in the right column. The results are not inconsistent with the multiple cluster scan statistic results, with both analyses giving very little evidence of brain cancer clusters, with the highest posterior probabilities of cluster membership being less than 0.006. In Figure 10 we present a barplot of the prior probabilities of clusters/anti-clusters for $j = 0,\ldots, 7$, along with the posteriors probabilities with and without adjustment for income. The posterior probability of no clusters/anti-clusters is approximately 0.99 in both cases, which is higher than the prior probability of $\pi_0 = 0.95$. In Figure 11 we present a plot of the sensitivity of the posterior probabilities on the number of clusters/anti-clusters as a function of the specification for π_0, the prior mass placed on the null configuration. The calculations for this plot can be carried out without re-running the analysis (for details, see Supplemental Materials). Tracing horizontal lines across the plot give the set of posterior probabilities for that π_0 value. There is a relative insensitivity to π_0 in both plots, with only very low (and unrealistic) values of π_0 giving even a single cluster/anti-cluster. In the Supplementary Materials we further investigate prior sensitivity to the specification of the narrow/wide components and an alternative prior on the number of clusters/anti-clusters (uniform on $1,\dots, J$). There is little sensitivity to the different priors, and for brain cancer we conclude there is little evidence for clusters in the Puget Sound region, at the geographical scale at which we have data available.

Bayesian Analysis for Breast Cancer

In Figure 12 we present the results for breast cancer using the income unadjusted expected counts in the left column and income adjusted expected counts in the right column. We observe a large number of areas with high posterior probability of cluster membership in Seattle and the Kirkland, Redmond, and Bellevue suburbs. These probabilities diminish in magnitude when using the income adjusted expected counts, indicating that a large component of the high relative risks in these areas has been absorbed by income. As we see in Figure 13, there is a large mass of posterior probability (94%) of $j = 4, 5$ clusters/anti-clusters when ignoring income information that shifts towards $j = 0$ once income information is incorporated. After adjustment there is 73% posterior probability on the null, with the majority of the remaining probability falling on 1 or 2 clusters/anti-clusters. There remains a single area in Tacoma with a relatively high posterior probability of 0.201. The second and third clusters indicated by the scan statistic method (Figure 7) do not have high posterior probabilities associated with them. The heat map in Figure 14 illustrates the sensitivity of the posterior on the number of clusters/anti-clusters to the prior on the null, π_0. In the case of breast cancer adjusted for income, the results are almost entirely insensitive to the choice of π_0. After incorporating income however, the posterior probabilities now vary more depending on the choice of π_0. Since higher values of π_0 correspond to higher prior skepticism at the existence of clusters/anti-clusters, the bulk of the posterior probabilities shifts towards $j = 0$. Additional summaries under different priors are given in the Supplementary Materials.

As discussed earlier, there are many well-documented risk factors for breast cancer, and we would expect many of these risk factors to have geographical structure. Unfortunately, as is usually the case in studies such as this based on aggregated count data, information on these risk factors is not available. Many of these risk factors will be associated with income, which explains why the number of clusters decreased when an area-level measure of income was included. This adjustment is clearly very crude and further analyses of breast cancer incidence in our study region would preferably be carried out using individual-level data with information on risk factors. Another important issue for breast cancer is screening. Screening rates vary geographically and we would expect those areas with high screening rates to have more breast cancer diagnoses even if the underlying risk is the same as areas with lower screening rates. This is supported by a plot of SMR versus screening rate (at the county level), included in the Supplementary Materials.

Discussion

Much like the multiple cluster Kulldorff method,[6] our Bayesian model constructs a set of single zones and treats each of them as potential clusters/anti-clusters. However, our method allows us to formally model the existence of more than one cluster by combining a number j of single zones to form configurations. These configurations, in particular the null configuration of no clusters/anti-clusters, are used as models to explain the observations and are the unknown parameters of our model to which we associate prior and posterior probabilities. By taking a Bayesian approach, our model accounts for sample size and power.[9;10;13]

While our method yields qualitatively similar results as the multiple cluster Kulldorff method, since our model allows for simultaneous modeling of more than one cluster/anti-cluster we obtain probabilities of cluster membership which are directly comparable between areas. This is in contrast to the multiple cluster Kulldorff method which considers only the most significant cluster, drops (if any) the most significant cluster, and iterates the procedure. The resulting p-values are difficult to interpret as they are sequential in nature and all resulting claims of significance based on them do not incorporate notions of sample size or power.

The computation for the Bayesian model is far greater than that for typical scan statistics, but the biggest difficulty in routinely using the Bayesian approach we have described is the need to specify prior distributions. In practice, as we have illustrated for brain and breast cancer, one should carry out sensitivity to the prior distribution. We would encourage the use of Kulldorff's scan statistic, in particular at the early stages of a cluster detection exercise, but we believe that the Bayesian approach we have described provides valuable additional information, foremost explicit posterior probabilities of cluster membership for each area, and a coherent way of handling multiple clusters.

The methods described in the paper is implemented in the R package `SpatialEpi`. An example of its application to the often studied upstate New York leukemia dataset can be found at https://github.com/rudeboybert/SpatialEpi.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Alexander, F.; Cuzick, J. Methods for the assessment of disease clusters. In: Elliott, P.; Cuzick, J.; English, D.; Stern, R., editors. Geographical and Environmental Epidemiology: Methods for Small-area Studies, pages. Oxford: Oxford University Press; 1992. p. 238-50.
2. Rothman KJ. A sobering start for the Cluster Buster's conference. Am J Epidemiol. 1990; 132(suppl):S6–13. [PubMed: 2356837]

3. Neutra RB. Counterpoint from a cluster buster. Am J Epidemiol. 1990; 132:1–8. [PubMed: 2356803]

4. Openshaw, S. The Modifiable Areal Unit Problem. Geo Books; Norwich: 1984. CATMOG No. 38

5. Besag J, Newell J. The detection of clusters in rare diseases. Journal of the Royal Statistical Society, Series A. 1991; 154:143–55.

6. Kulldorff M. A spatial scan statistic. Communications in Statistics: Theory and Methods. 1997; 26:1481–1496.

7. Kulldorff, M.; Rand, K.; Gherman, G.; Williams, G.; Francesco, D. SaTScan – software for the spatial and space-time scan statistics, version 9.3. National Cancer Institute; Bethedsa, Madison: 1998.

8. Savitz DA. Commentary: Reconciling theory and practice: What is to be done with P values? Epidemiology. 2013; 24(2):212–214. [PubMed: 23377090]

9. Greenland S, Poole C. Rejoinder: Living with statistics in observational research. Epidemiology. 2013; 24(1):73–78. [PubMed: 23232613]

10. Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. American Journal of Epidemiology. 1993; 137(5):485496.

11. Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast cancer clusters in the northeast united states: a geographic analysis. American Journal of Epidemiology. 1997; 146:161–170. [PubMed: 9230778]

12. Jemal A, Kulldorff M, Devesa SS, Hayes RB, Fraumeni JF. A geographic analysis of prostate cancer mortality in the united states, 1970–89. International Journal of Cancer. 2002; 101:168–174. [PubMed: 12209994]

13. Wakefield J. Bayes factors for genome-wide association studies: comparison with p-values. Genetic Epidemiology. 2009; 33(1):79–86. [PubMed: 18642345]

14. Zhang Z, Assunção R, Kulldorff M. Spatial scan statistics adjusted for multiple clusters. Journal of Probability and Statistics. 2010 page Article ID 642379.

15. Wakefield, JC.; Best, NG.; Waller, LA. Bayesian approaches to disease mapping. In: Elliott, P.; Wakefield, JC.; Best, NG.; Briggs, D., editors. Spatial Epidemiology: Methods and Applications. Oxford University Press; Oxford: 2000. p. 104-27.

16. Richardson S, Thomson A, Best NG, Elliott P. Mini-monograph: Interpreting posterior relative risk estimates in disease mapping studies. Environmental Health Perspectives. 2004; 112:1016–1025. [PubMed: 15198922]

17. Wakefield JC, Kim A. A Bayesian model for cluster detection. Biostatistics. 2013; 14:752–765. [PubMed: 23476026]

18. Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC. Monitoring for clusters of disease: application to leukaemia incidence in upstate New York. American Journal of Epidemiology. 1990; 132:S136–S143. [PubMed: 2356825]

19. Devesa, SS.; Grauman, DJ.; Blot, WJ.; Hoover, RN.; Fraumeni, JF. Atlas of Cancer Mortality in the United States 1950–94. National Institutes of Health; 1999. NIH Publications No. 99–4564

20. Fang Z, Kulldorff M, Gregorio DI. Brain cancer in the United States, 1986–95: A geographic analysis. Neuro-Oncology. 2003; 6:179–187. [PubMed: 15279710]

21. Gangnon RE, Clayton MK. Bayesian detection and modeling of spatial disease clustering. Biometrics. 2000; 56:922–935. [PubMed: 10985238]

22. Gangnon RE, Clayton MK. A hierarchical model for spatially clustered disease rates. Statistics in Medicine. 2003; 22:3213–3228. [PubMed: 14518024]

23. Gangnon RE. Impact of prior choice on local Bayes factors for cluster detection. Statistics in Medicine. 2006; 25:883–895. [PubMed: 16453368]

24. Ries, LAG.; Melbert, D.; Krapcho, M.; Stinchcomb, DG.; Howlader, N.; Horner, NJ.; Mariotto, A.; Miller, BA.; Feuer, EJ.; Altekruse, SF.; Lewis, DR.; Clegg L, L.; Eisner, MP.; Reichman, M.; Edwards, BK., editors. SEER Cancer Statistics Review, 1975–2005. National Cancer Institute; Bethesda MD: 2008.

25. Cancer surveillance system. http://www.fredhutch.org/en/labs/phs/projects/cancer-surveillance-system.html. Accessed: 2015-02-08

26. Jerks GF. The data model concept in statistical mapping. International Yearbook of Cartography. 1967; 7:186–190.

27. Joseph Sheehan T, DeChello Laurie M, Kulldorff Martin, Gregorio David I, Gershman Susan, Mroszczyk Mary. The geographic distribution of breast cancer incidence in massachusetts 1988 to 1997, adjusted for covariates. International Journal of Health Geographics. 2004; 3(1):17. [PubMed: 15291960]

28. Ed Hsu, Chiehwen; Jacobson, Holly; Soto Mas, Francisco. Evaluating the disparity of female breast cancer mortality among racial groups – A spatiotemporal analysis. International Journal of Health Geographics. 2004; 3(1):4. [PubMed: 14987336]

29. Gregorio, David I.; Kulldorff, Martin; Barry, Leah; Samociuk, Holly. Geographic differences in invasive and in situ breast cancer incidence according to precise geographic coordinates, connecticut, 1991–95. International journal of cancer. 2002; 100(2):194–198. [PubMed: 12115569]

30. Tian, Nancy; Gaines Wilson, J.; Benjamin Zhan, F. Female breast cancer mortality clusters within racial groups in the united states. Health & Place. 2010; 16(2):209–218. [PubMed: 19879177]

31. Sariego, Jack. Patterns of breast cancer presentation in the united states: does geography matter? The American Surgeon. 2009; 75(7):545–550. [PubMed: 19655596]

32. Jacquez, Geoffrey M.; Greiling, Dunrie A. Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxics in Long Island, New York. International Journal of Health Geographics. 2003; 2:4–25. [PubMed: 12633502]

33. Fukuda Y, Umezaki M, Nakamura K, Takano T. Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan. Cancer Causes & Control. 2006; 17:449–457. [PubMed: 16596297]

34. Elliott, P.; Wakefield, JC.; Best, NG.; Briggs, DJ. Spatial Epidemiology: Methods and Applications. Oxford University Press; Oxford: 2000.

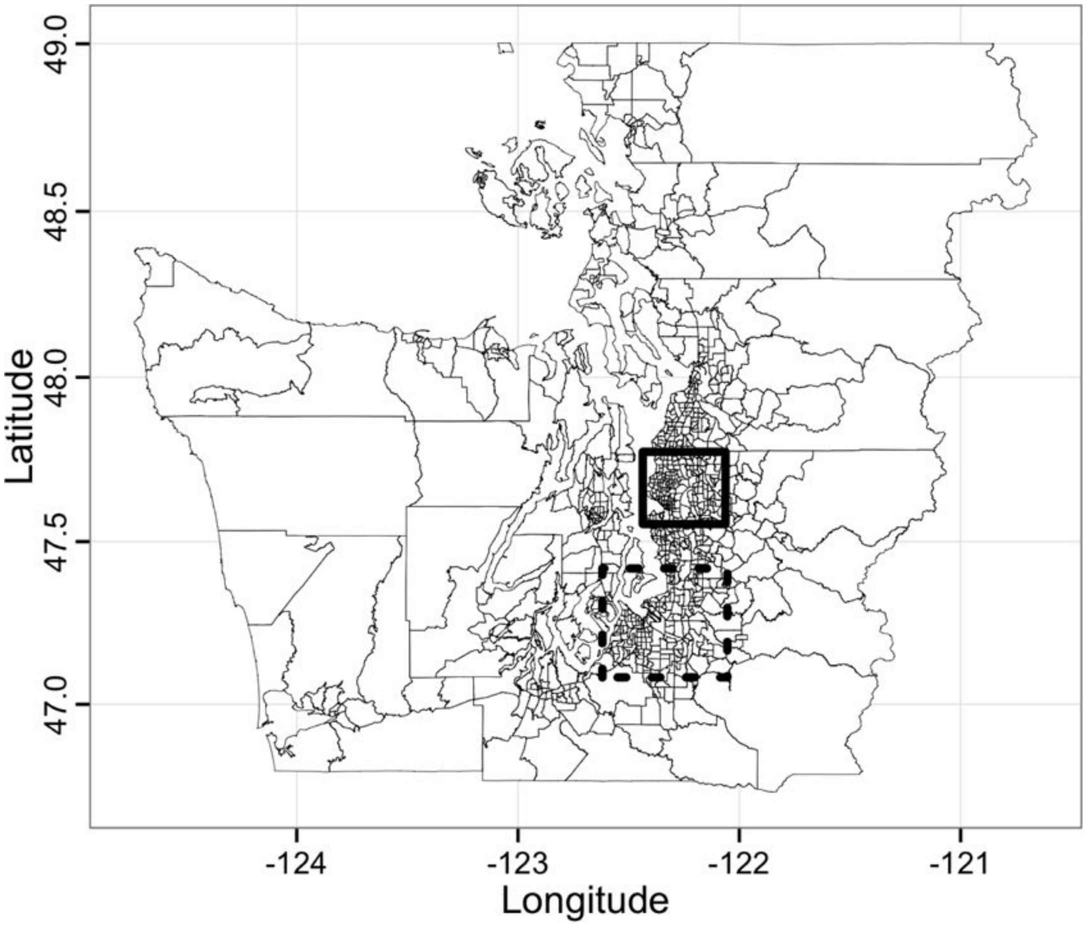

Figure 1.

Study region with Seattle (solid line) and Tacoma (dashed line) metropolitan regions highlighted and the city of Mount Vernon marked with a dot.

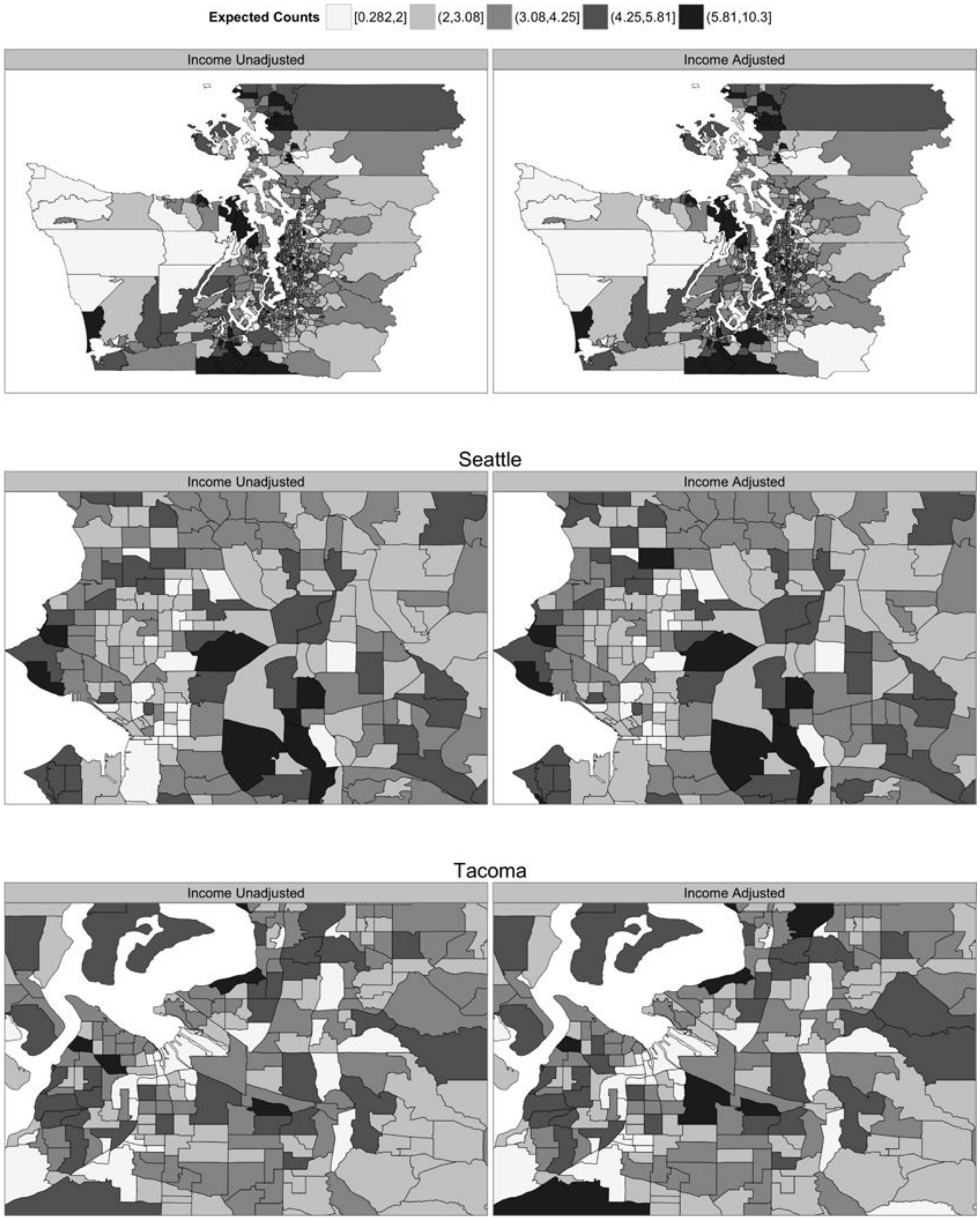

Figure 2.

Maps of income unadjusted (left column) and income adjusted (right column) expected counts of brain cancer.

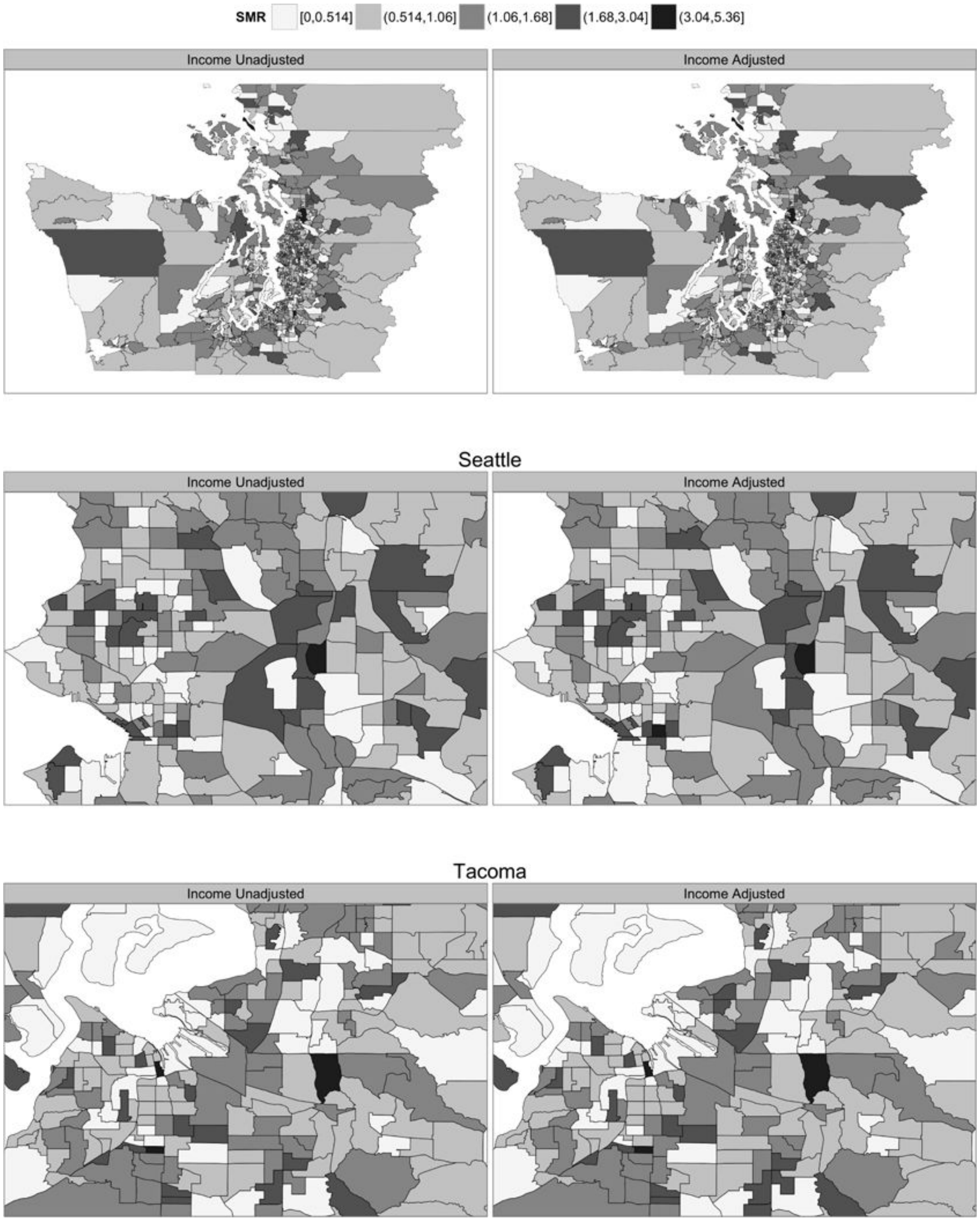

Figure 3.

Maps of income unadjusted (left column) and income adjusted (right column) standardized morbidity ratios for brain cancer.

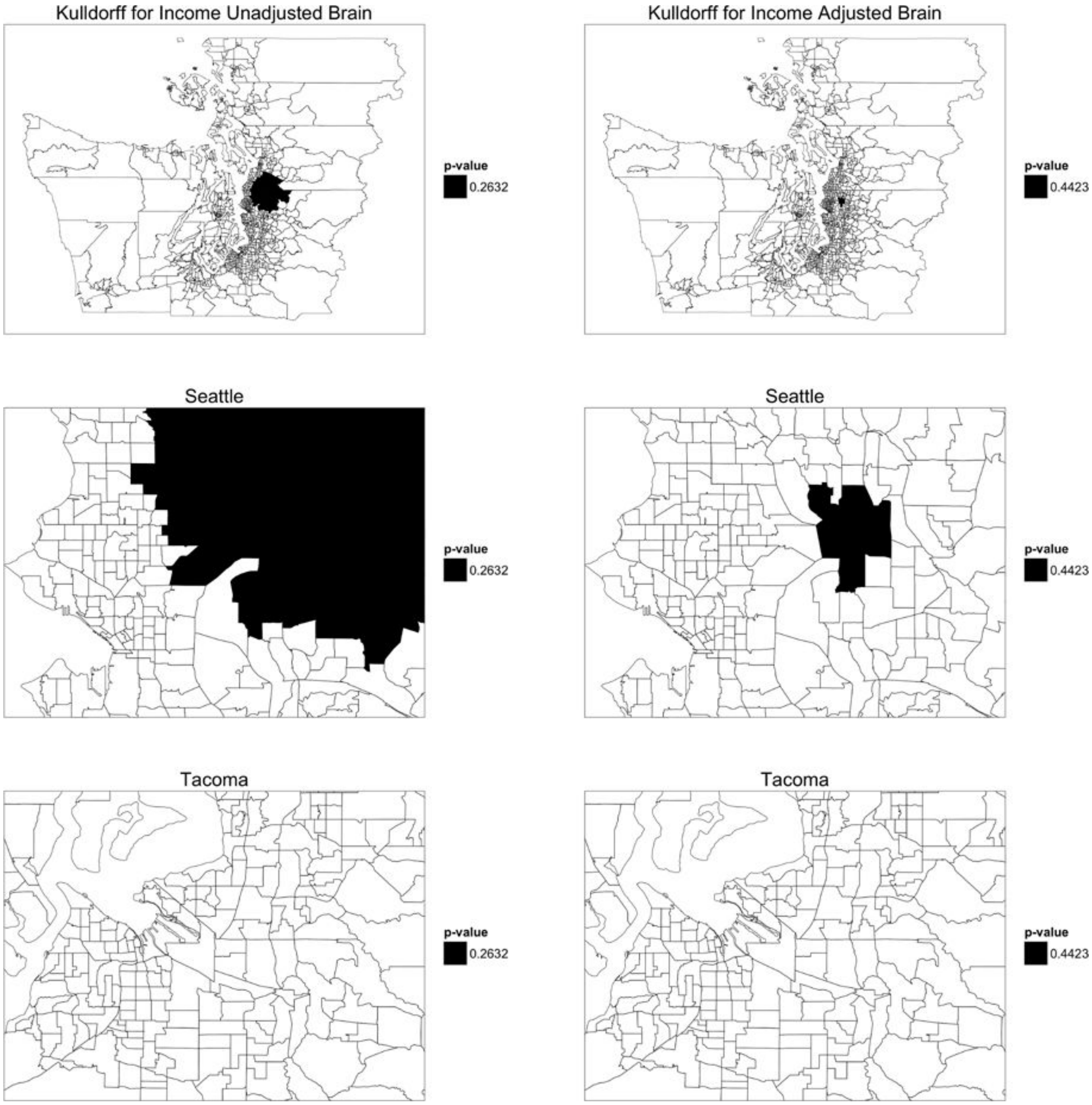

Figure 4.

Maps of income unadjusted (left column) and income adjusted (right column) multiple cluster scan statistic results for brain cancer.

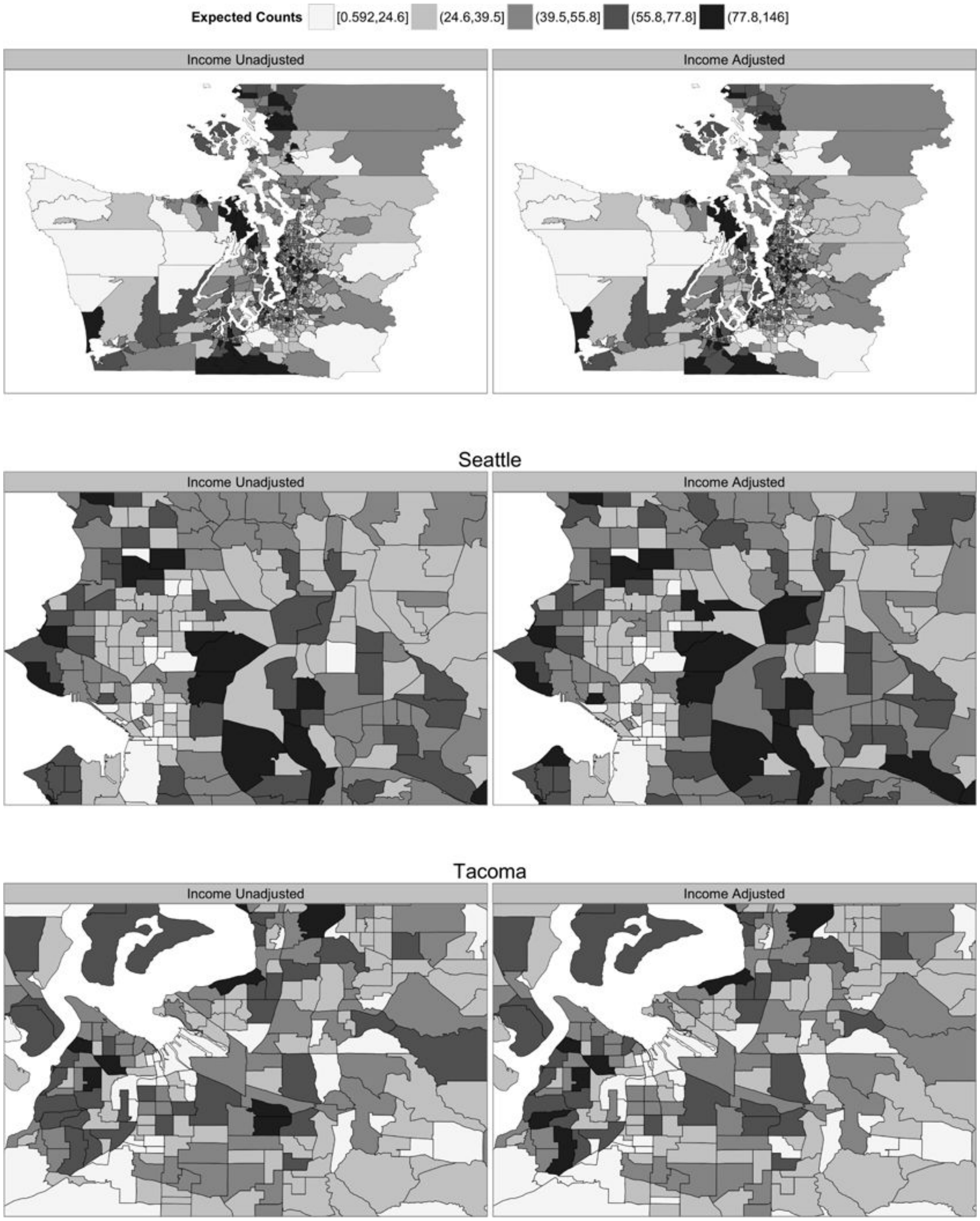

Figure 5.

Maps of income unadjusted (left column) and income adjusted (right column) expected counts of breast cancer.

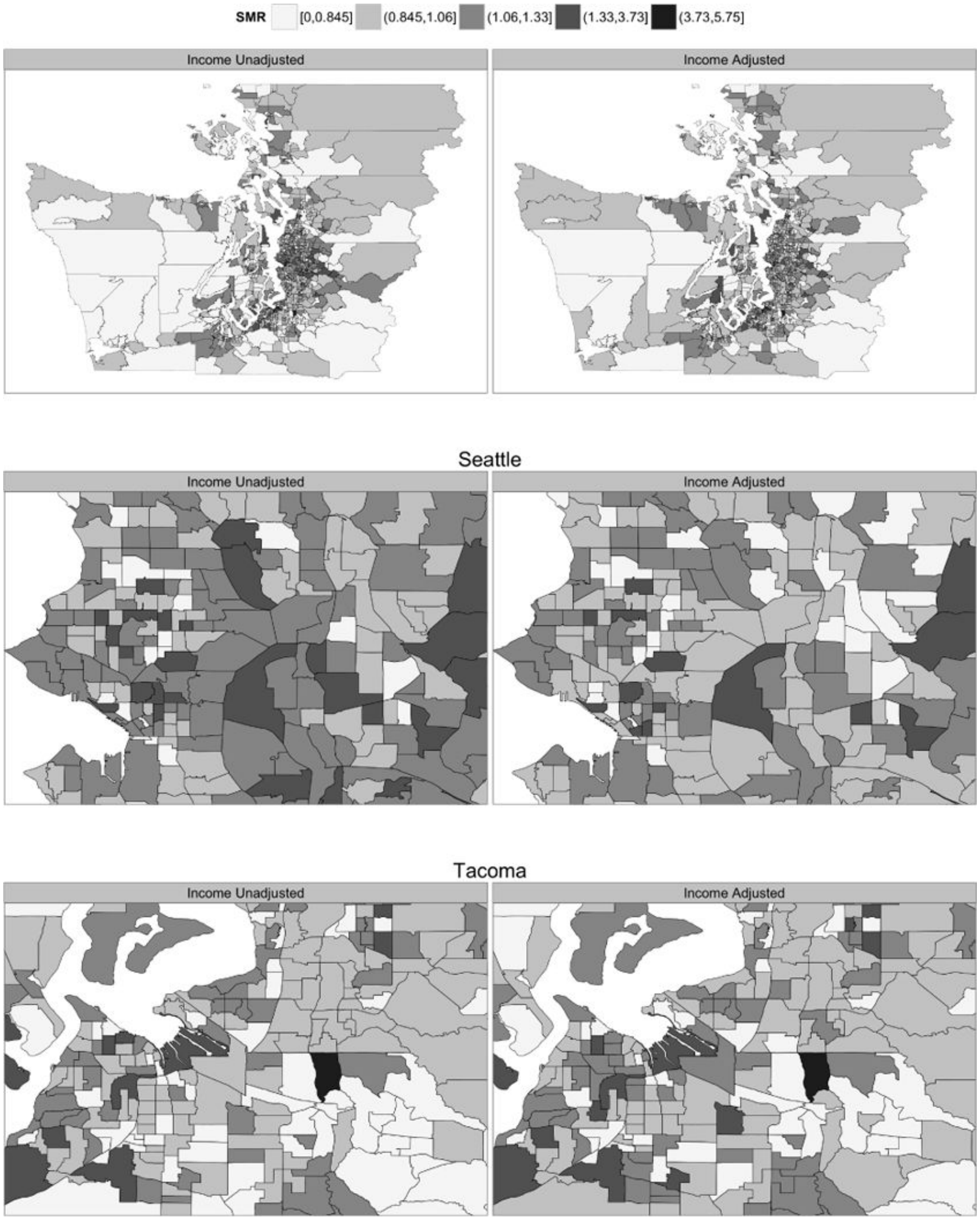

Figure 6.

Maps of income unadjusted (left column) and income adjusted (right column) standardized morbidity ratios for breast cancer.

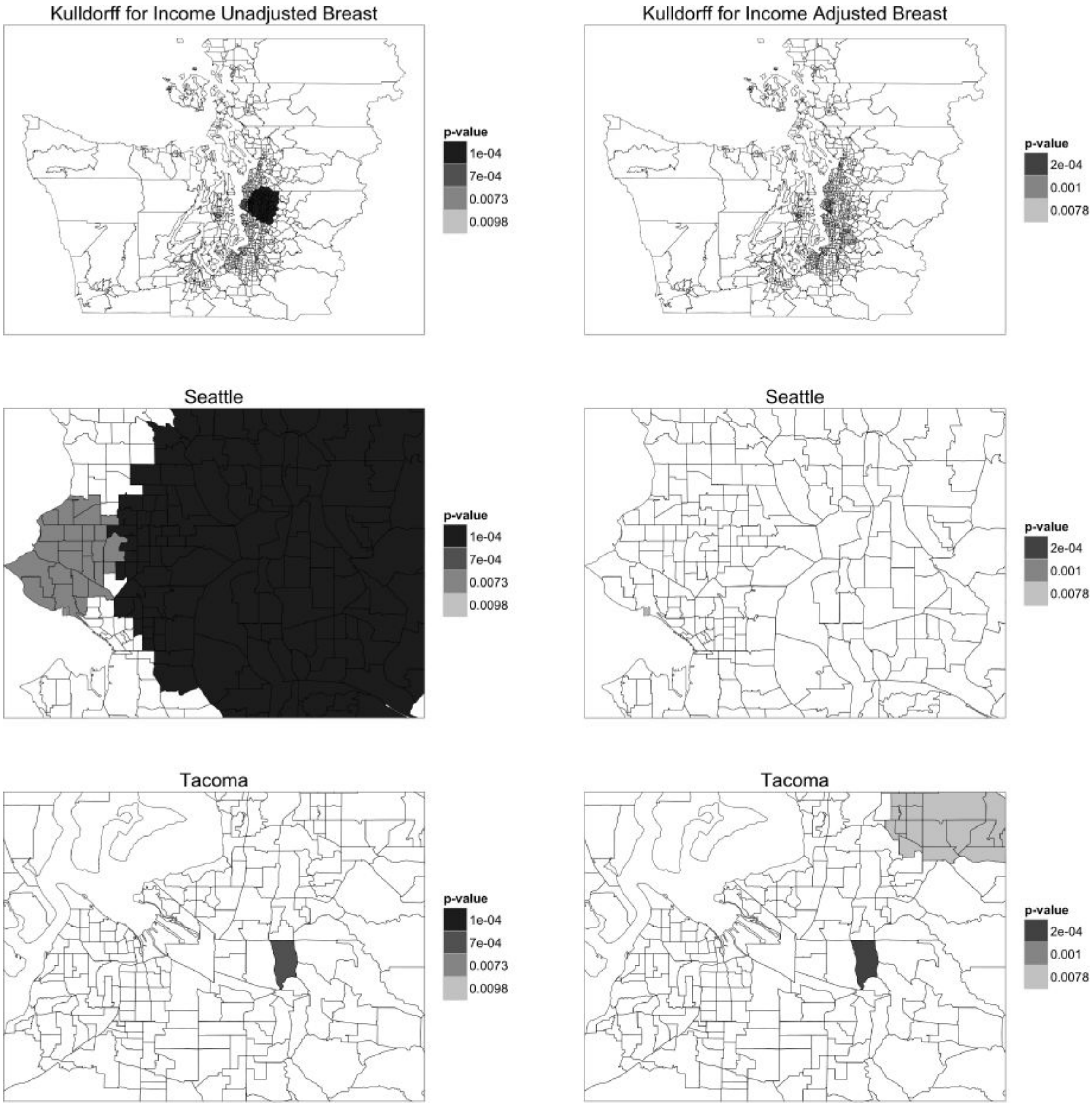

Figure 7.

Maps of income unadjusted (left column) and income adjusted (right column) multiple cluster scan statistic results for breast cancer.

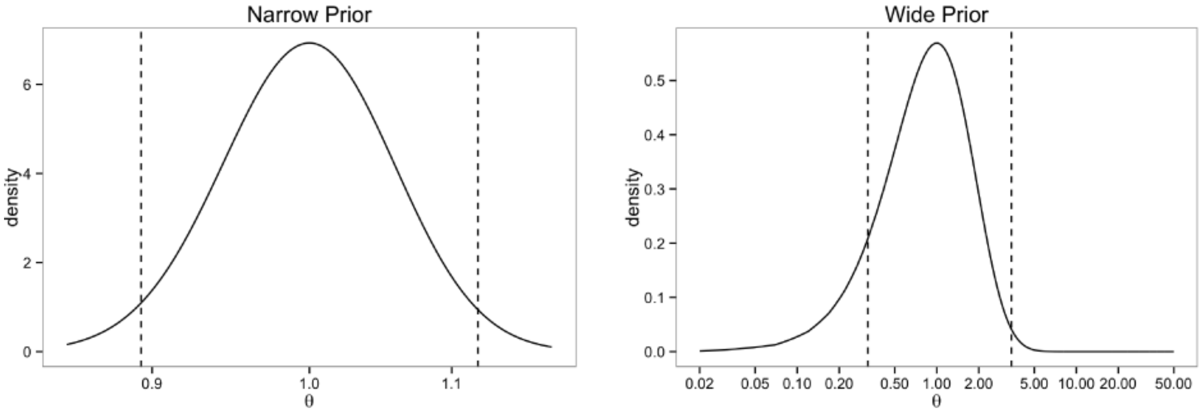

Figure 8.

Wide and narrow distributions on the relative risk. Under the null (no clusters/anti-clusters), relative risks θ are assumed to arise from the narrow prior, so that there is still a small amount of "wobble" about 1. Under the alternative (at least one cluster/anti-cluster), the relative risks are assumed to arise from the wide prior, so that there is greater variation. Note that the θ scale is logarithmic.

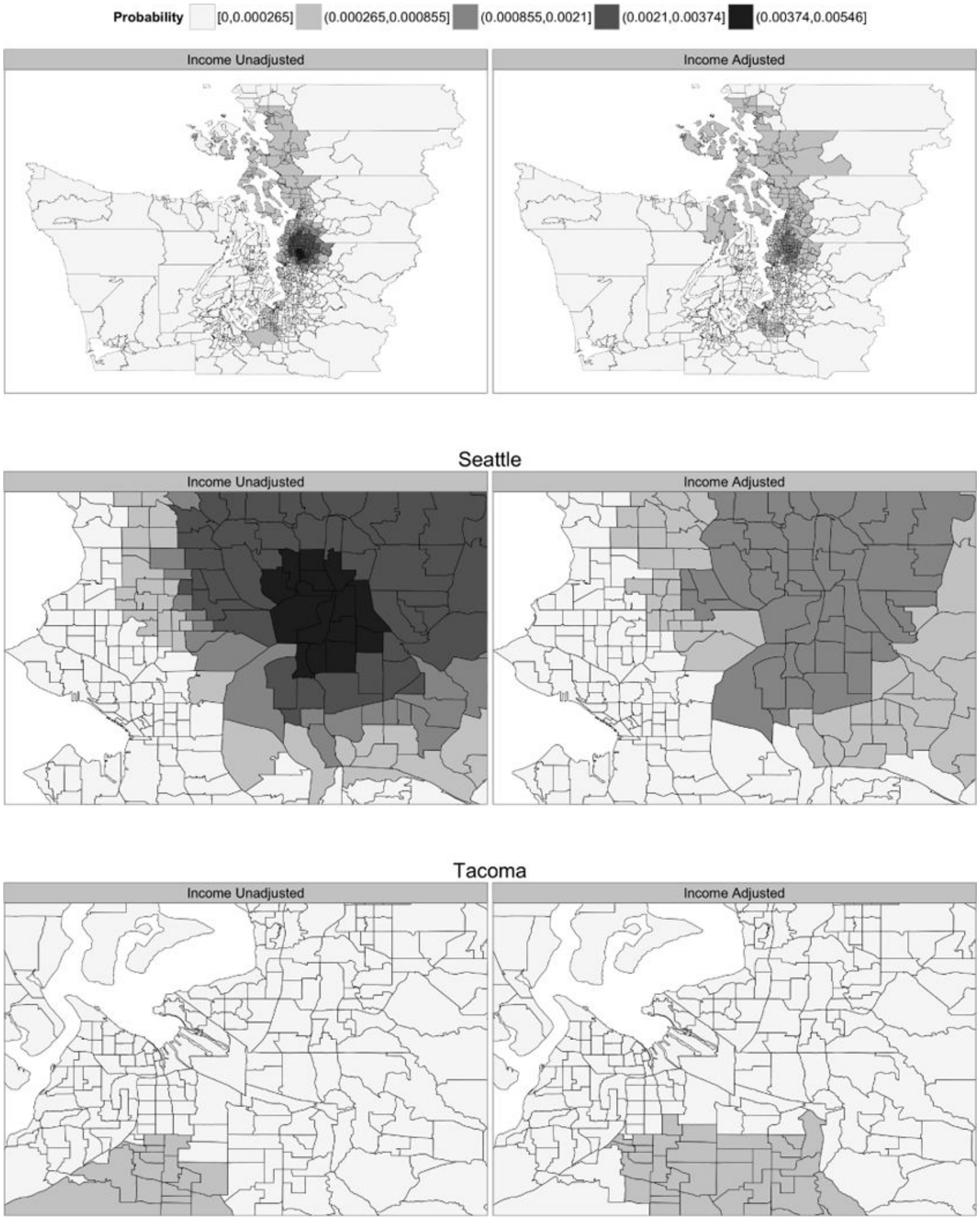

Figure 9.

Maps of income unadjusted (left column) and income adjusted (right column) posterior probabilities of brain cancer cluster membership.

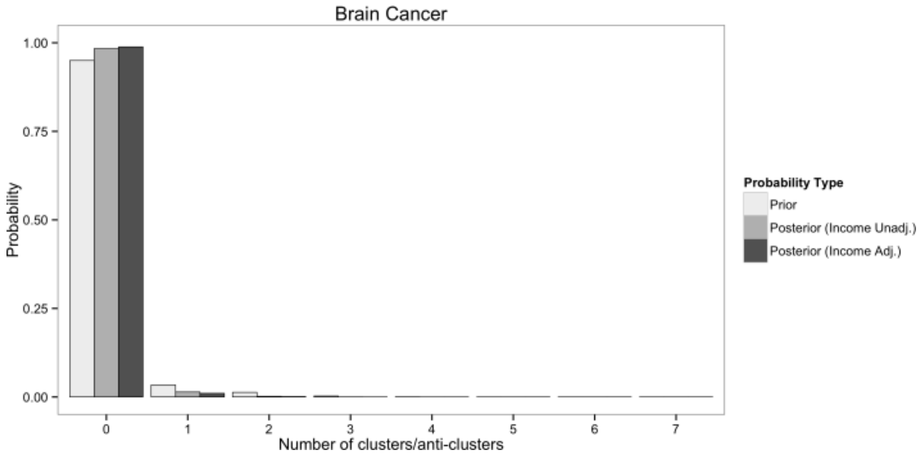

Figure 10.

Prior/posterior probabilities of the number of brain cancer clusters/anti-clusters.

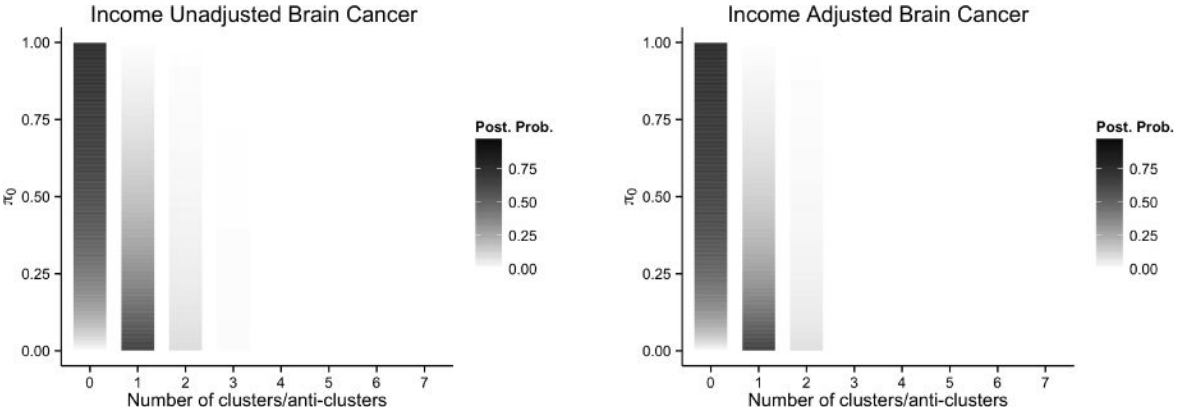

Figure 11.

Sensitivity of posterior probabilities of the number of brain cancer clusters/anti-clusters to π_0. Tracing horizontal lines across the plot give the set of posterior probabilities for that π_0 value.

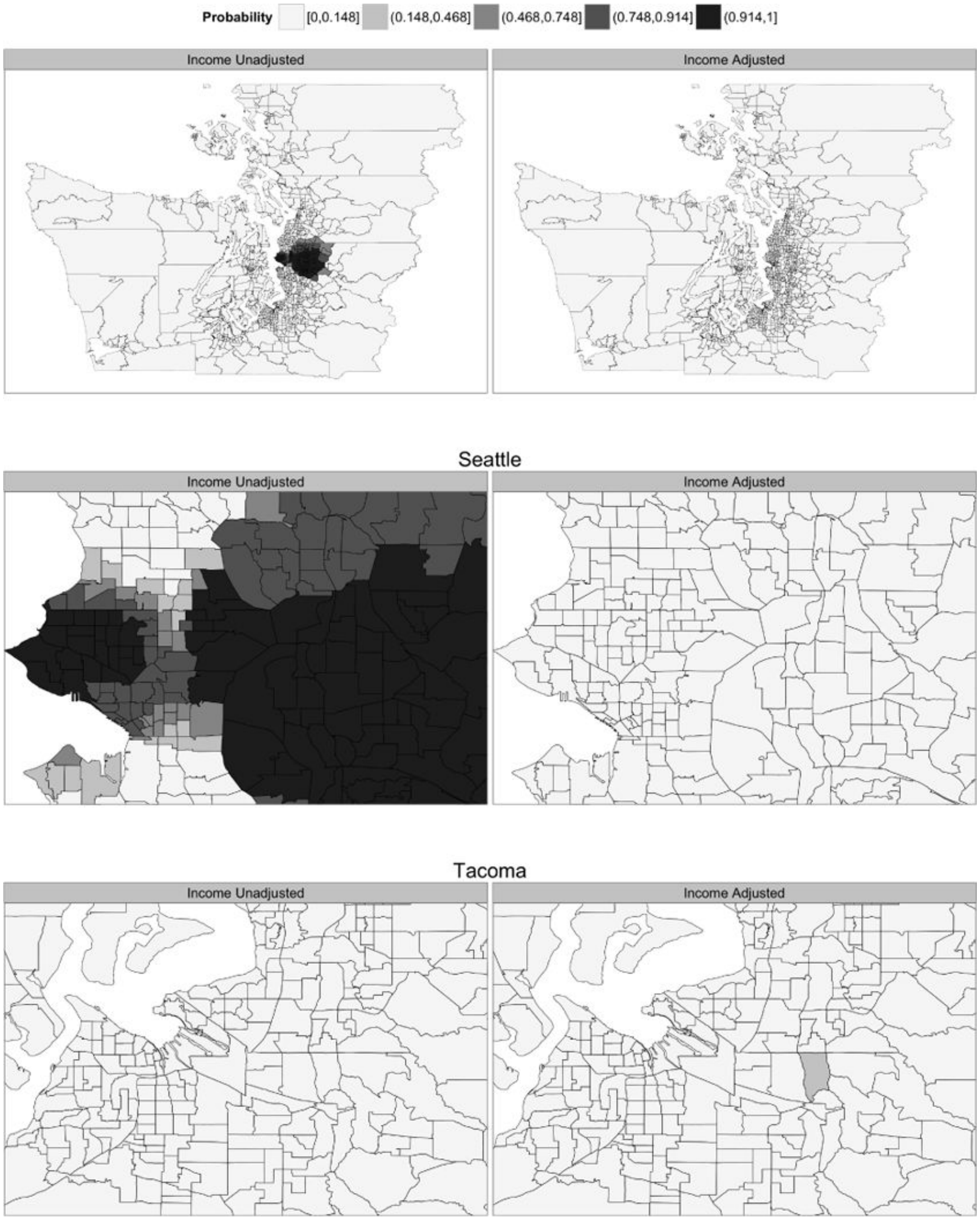

Figure 12.

Maps of income unadjusted (left column) and income adjusted (right column) posterior probabilities of breast cancer cluster membership.

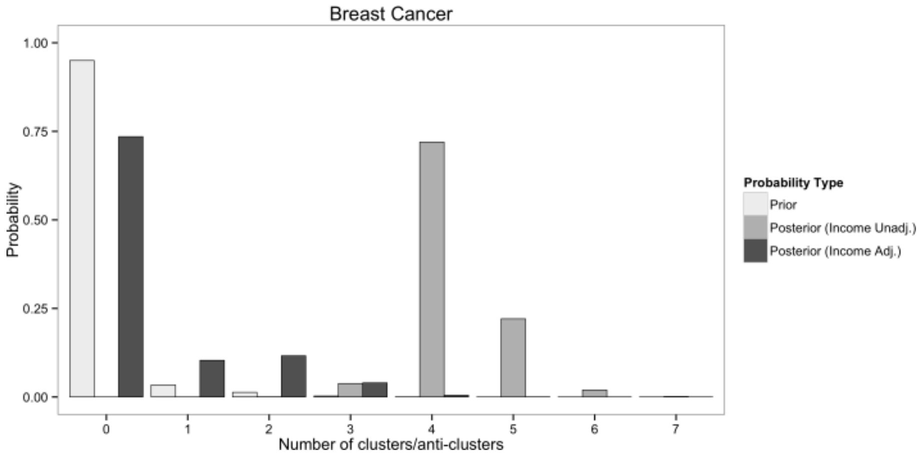

Figure 13.

Prior/posterior probabilities of the number of breast cancer clusters/anti-clusters.

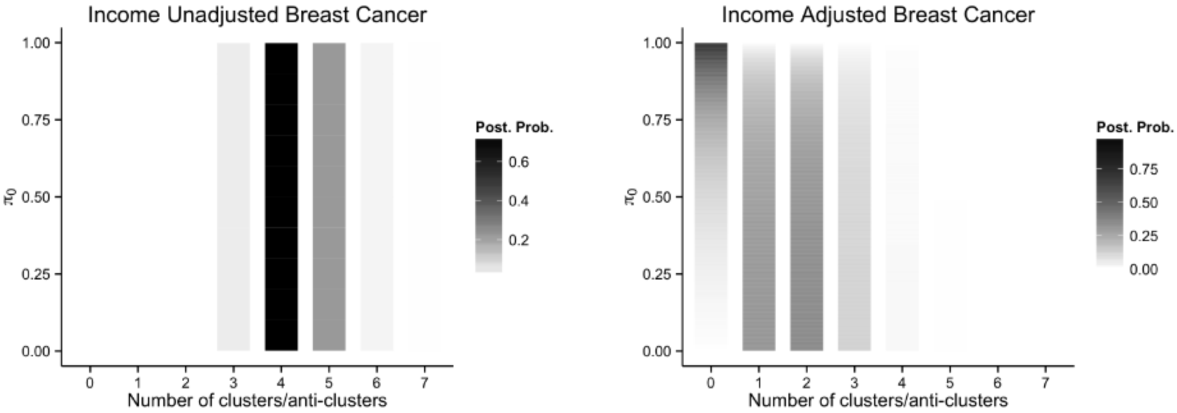

Figure 14.

Sensitivity of posterior probabilities of the number of breast cancer clusters/anti-clusters to π_0. Tracing horizontal lines across the plot give the set of posterior probabilities for that π_0 value.

Table 1

Summary statistics for five cancer sites across the study region. Expected counts E and standardized morbidity ratios SMR are not adjusted for income, while E* and SMR* are income adjusted. The total number of cases for brain, (females) breast, lung, colorectal, and (male) prostate cancers are: 2935, 37726, 26392, 19195, and 29117.

Type	Value	Min	5th %ile	Median	95th %ile	Max
	Disease Counts y	0.00	0.00	3.00	7.00	16.00
	Expected Counts E	0.30	1.42	3.15	5.68	9.95
Brain	SMR	0.00	0.00	0.95	2.09	5.24
	Expected Counts E*	0.28	1.36	3.14	5.81	10.30
	SMR*	0.00	0.00	0.97	2.11	5.36
	Disease Counts y	0.00	13.00	39.00	82.70	124.00
	Expected Counts E	0.69	16.07	39.81	79.48	131.20
Breast	SMR	0.00	0.67	0.99	1.38	5.11
	Expected Counts E*	0.59	15.32	39.51	81.19	145.67
	SMR*	0.00	0.69	1.00	1.36	5.75
	Disease Counts y	0.00	9.00	27.00	61.00	101.00
	Expected Counts E	0.26	10.44	27.15	60.33	105.23
Lung	SMR	0.00	0.53	1.00	1.70	3.86
	Expected Counts E*	0.36	10.26	27.03	59.85	106.90
	SMR*	0.00	0.60	1.00	1.57	2.84
	Disease Counts y	0.00	6.00	20.00	43.00	67.00
	Expected Counts E	0.45	7.57	19.82	44.25	74.22
Colorectal	SMR	0.00	0.57	1.00	1.60	2.77
	Expected Counts E*	0.47	7.63	19.64	43.41	68.52
	SMR*	0.00	0.58	0.99	1.60	3.05
	Disease Counts y	0.00	10.00	29.00	67.00	132.00
	Expected Counts E	0.17	12.01	30.11	62.50	108.85
Prostate	SMR	0.00	0.63	0.98	1.44	4.82
	Expected Counts E*	0.15	11.66	29.51	64.13	109.35

Type	Value	Min	5th %ile	Median	95th %ile	Max
	SMR*	0.00	0.65	0.97	1.44	5.19