Journal of Data Science **6**(2008), 105-123

Joint Spatio-Temporal Modeling of Low Incidence Cancers Sharing Common Risk Factors

Jacob J. Oleson[1], Brian J. Smith[1] and Hoon Kim[2]
[1] *The University of Iowa and* [2] *California State Polytechnic University*

Abstract: In this article, we present a joint modeling approach that combines information from multiple diseases. Our model can be used to obtain more reliable estimates in rare diseases by incorporating information from more common diseases for which there exists a shared set of important risk factors. Information is shared through both a latent spatial process and a latent temporal process. We develop a fully Bayesian hierarchical implementation of our spatio-temporal model in order to estimate relative risk, adjusted for age and gender, at the county level in Iowa in five-year intervals for the period 1973–2002. Our analysis includes lung, oral, and esophageal cancers which are related to excessive tobacco and alcohol use risk factors. Lung cancer risk estimates tend to be stable due to the large number of occurrences in small regions, i.e. counties. The lower risk cancers (oral and esophageal) have fewer occurrences in small regions and thus have estimates that are highly variable and unreliable. Estimates from individual and joint modeling of these diseases are examined and compared. The joint modeling approach has a profound impact on estimates regarding the low risk oral and esophageal cancers while the higher risk lung cancer is minutely impacted. Clearer spatial and temporal patterns are obtained and the standard errors of the estimates are reduced leading to more reliable estimates.

Key words: Autoregressive prior, CAR, disease mapping, hierarchical Bayes, Markov chain Monte Carlo.

1. Introduction

The spatial analysis of disease incidence data, known as disease mapping, is a dynamic area of biostatistical, epidemiological, and public health research. Using geographical mapping, we can detect "hot-spots" of disease incidence in which nearby areas are often related because they share similar risk factors. This attention has led to a greater use of geographical, GIS, and spatial analysis tools in studying data routinely collected for public health purposes.

Disease mapping is commonly used to describe the variation in health outcomes over geographic regions. Mapping of crude disease rates can be quite

misleading, particularly at a small area (county) level. This is often due to the relatively small incidence counts in regions and the presence of spatial correlation in the rates. With large amounts of data at the state level, estimates are quite reliable. At the small area level though, the number of incident cases tends to be much smaller resulting in unreliable estimates. High prevalence diseases may have a large amount of information at the county level. For low prevalence diseases, however, stable estimates at the county-specific level are difficult to attain. A statistical model that combines information from related diseases can reduce the variability in estimates and help in identifying "hot-spots" for less prevalent diseases which, in turn, improves potential prediction of the diseases.

Most methods for disease mapping focus on the spatial modeling of a single disease. Some recent developments have examined relationships between multiple diseases (Knor-Held and Best, 2001; Kim *et al.* 2001; and Held *et al.*, 2005). Held *et al.* (2005) also consider a Bayesian shared component model. That model, however, ignores any temporal changes. Kim *et al.* (2001) proposes a Bayesian joint spatio-temporal analysis but in a framework limited to two diseases. In this article, we proposed a framework that can easily incorporate any number of diseases while accounting for temporal correlation. A Bayesian hierarchical modeling approach is taken in the development of our new technique for relating two or more diseases.

Our research is motivated by an investigation of the spatial and temporal variation in lung, oral, and esophageal cancer rates at the county level in the state of Iowa for the years 1973–2002. These cancers are all highly associated with tobacco-related risk factor and all but lung cancer are further associated with excessive alcohol consumption. There may be an indirect relationship between lung cancer and excessive alcohol consumption though, in that most people who smoke also drink alcohol. Because of the inherent relationships between these cancers, we take a joint modeling approach to estimate and map their corresponding relative risks. Patterns for less frequent cancers (oral and esophageal) are difficult to visualize across space due to high variability as a result of the small number of cases. We expect to reduce that variability by incorporating information from a more prevalent disease (lung) which shares similar risk factors both spatially and temporally.

From the disease mapping, we can detect "hot-spots" of disease incidence in which nearby geographic areas are often related because they share similar risk factors. When diseases share common risk factors, such as tobacco and alcohol use which are more prevalent in some communities than in others, the correlation between diseases leads to similar geographic and temporal patterns. In such settings, information from one or more diseases can help estimate relative risks for another disease. This is particularly helpful when data are unavail-

able for the risk factors and the disease of interest has a very low incidence. The additional information "borrowed" from the related diseases can help reduce variability in the relative risk estimate, thus strengthening the estimate based on low counts. This relationship can be recognized through a joint analysis. We develop a Bayesian hierarchical model to combine information from multiple diseases that share common risk factors in order to facilitate the mapping and elucidation of spatio-temporal patterns in disease incidence.

We outline the data and motivation of the problem in Section 2. We then describe the hierarchical model and Bayesian formulation in Section 3. In Section 4, the modeling results are presented. This begins with a look at modeling estimates obtained for each of the three cancers without any joint information and concludes with presenting the estimates using the joint model. Estimates from the two methods are compared. We conclude with a discussion in Section 5.

2. Data and Motivation

Smoking and excessive alcohol use are risk factors for a large number of cancers. According to the National Cancer Institute (www.cancer.gov), smoking damages nearly every organ in the body and is linked to at least ten different cancers. It accounts for nearly 30% of all cancer deaths—a primary reason why tobacco use is the leading modifiable risk factor for cancer. Tobacco use is specifically associated with cancers of the lung and bronchus, oral cavity (excluding lip), and esophagus. Of these tobacco-related cancers, the strongest associations between alcohol use and cancer are in cancers of the oral cavity and esophagus. The risk of these cancers increases significantly when tobacco and alcohol are used together.

The necessary data for our analysis are available through the National Cancer Institute's SEER*Stat program[1]. According to SEER data, incident lung cancers were observed at a rate of 72.2 cases per 100,000 individuals in Iowa during 2002. Iowa lung cancer rates have increased in recent years. In particular, statewide incidence rates increased by 60% during the period 1973–2002. During this time, rates in Iowa went from being much lower than the national average (41.5 per 100,000 people in 1973) to being higher than the national average (58.9 per 100,000 people in 2002). The increase was more profound in some counties than others. We note that incidence rates in Iowa's metropolitan statistical area (MSA) counties peaked around 1984 and have subsequently shown a decline. Rural counties in Iowa, however, appear to continue a general increase in incidence

[1]See www.seer.cancer.gov (2006). Surveillance, Epidemiology, and End Results (SEER) Program, Public-Use Data (1973-2002), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch.

rates and mortality rates. Similar increases can be seen for oral and esophageal cancers for which the respective incidence rate increases went from 7.6 to 9.9 per 100,000 and 2.1 to 5.6 per 100,000 individuals during 1973–2002. As is the case for lung cancer, Iowa's rates for these cancers are well below the 1973 national averages of 10.1 per 100,000 and 4.5 per 100,000 yet above the 2002 national averages of 9.8 per 100,000 and 4.4 per 100,000 for oral and esophageal cancers, respectively.

Changes over time are more difficult to detect at the county-specific level for cancers that have low incidence rates such as oral and esophageal cancers. County-specific estimates based on low counts tend to be more variable and unreliable. For example, the risk of oral cancer in Iowa for 2002 is an estimated 30% higher than the risk in 1973. Since oral cancer incidence is about 7 times less than lung cancer incidence, the associated risk estimates are considerably less precise. As noted previously for lung cancer, the temporal trend for oral cancer appears to be more profound in some counties than others. Moreover, the county-to-county variation appears to be greater for oral cancer, but at least part of that perception is an artifact due to extremely small counts.

Typically, incidence rates are calculated for regions where population counts are readily available (e.g., per county). It is also useful to compare the observed number of cases in a region to the number that would be expected in the at-risk population in order to adjust for important demographic covariates. For instance, we apply the national crude incidence rate to the county-specific age and gender distributions to obtain expected counts for each of the three cancers under our investigation. The ratio of observed to expected counts provides an estimate of the Standardized Morbidity Ratio (SMR). This ratio is an estimate of the age- and gender-adjusted relative risk in each county. In this study, we will focus on incidence, rather than mortality, as our disease outcome of interest and use a fully Bayesian approach to estimate the SMR.

Relative risks are not confined to political boundaries and the underlying risk factors are also not constrained to these boundaries. It may be reasonable to assume that relative risks in neighboring counties are generally similar (correlated) through a spatial process that could be the result of a number of factors; e.g., environmental condition, work-related risk factors, or lifestyle choices. This underlying spatial process is likely to be very similar for those cancers that share common or similar risk factors.

In addition to being correlated over space, the diseases may follow related patterns over time due to temporal changes in the shared risk factors. Thus, our model also includes a joint temporal relationship. Because of the small number of occurrences for these cancers on a yearly basis, we aggregate into six intervals of five year increments. They are 1973–1977, 1978–1982, 1983–1987, 1988–1992,

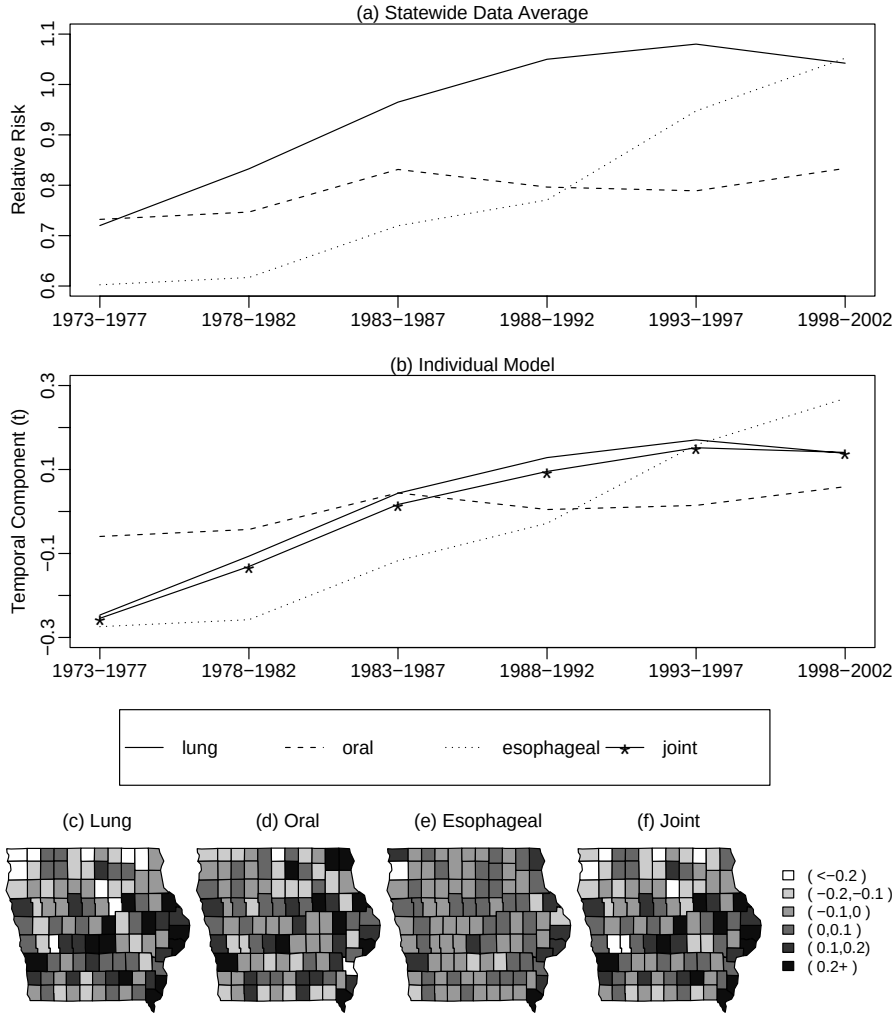

Figure 1: Modeling temporal and spatial effects. (a) MLEs of the state-wide relative risks for lung, oral, and esophageal cancers. (b) Posterior means for the latent temporal parameter t in the individual and joint Bayesian models. The following plot means for the latent spatial process, Z_i, in the individual disease models: (c) Lung Cancer, (d) Oral Cancer, and (e) Esophageal Cancer. (f) The posterior mean spatial process in the joint model.

1993–1997, and 1998–2002. Many methods of correlating information over time assume data are stationary. This is a safe assumption to make for the three cancers in our study. While the incidence rates and relative risks have shown

an increase since 1973, there is reason to believe that those rates have leveled off and will begin to decrease in the future (See Figure 1(a)). According to the Behavioral Risk Factor Surveillance System (BRFSS), binge drinking among Iowa adults has shown an increase from 13.4% in 1990 to 20.1% in 2002. These numbers are increasing faster than the U.S. median which only rose from 15.3% to 16.1% during that time span. Smoking levels reported by the BRFSS in Iowa have increased from 21.7% in 1990 to 23.2% in 2002. Levels peaked in 1996 at 23.6%. The nationwide median has remained steady at 23.0% over this same time frame.

3. Hierarchical Model

3.1 Linear model

We propose a hierarchical fully Bayesian model to provide improved estimates of county-specific cancer relative risks. Let Y_{ijk} be the number of cancer incidences, where i represents the county (1–99), j the six time periods, and k the cancer (lung, oral, and esophageal). Given the SMR θ_{ijk}, the number of incidences, Y_{ijk}, for each county, time, and cancer category is assumed to follow a Poisson distribution with mean $E_{ijk}\theta_{ijk}$, where E_{ijk} denotes the expected number of cases adjusted for age and gender and is a fixed quantity in the model. It is calculated by applying the national crude incidence rate for the associated cancer to the county-specific age and gender distributions (available through the SEER*Stat program). The parameter θ_{ijk} represents the true, but unknown SMR. A ratio less than one indicates risk less than the national average whereas a ratio greater than one indicates risk greater than the national average. The Poisson distributional assumption for Y_{ijk} is appropriate for rare and non-contagious diseases.

We will employ a log-linear model for the cancer incidence rates. By including effects for spatial and temporal variation, our model will account for variability in θ_{ijk} due to county, year, and cancer. We propose a latent spatial process representing relationships among neighboring counties, similar to that of Oleson and He (2006). We also designate a latent temporal process to characterize correlation over time. The underlying spatio-temporal relationship is driven by the common tobacco and alcohol risk factors. The model is specified using the following:

$$\begin{aligned} Y_{ijk} &\sim Poisson\left(E_{ijk}\theta_{ijk}\right) \\ \log(\theta_{ijk}) &= \alpha_k + \phi_k Z_i + \psi_k t_j + e_{ijk}. \end{aligned} \tag{3.1}$$

In this framework, $i = 1, \ldots, 99$ are the counties, $j = 1, \ldots, 6$ are the time periods, and $k = 1, 2, 3$ are lung, oral and esophageal cancers, respectively. The parameter

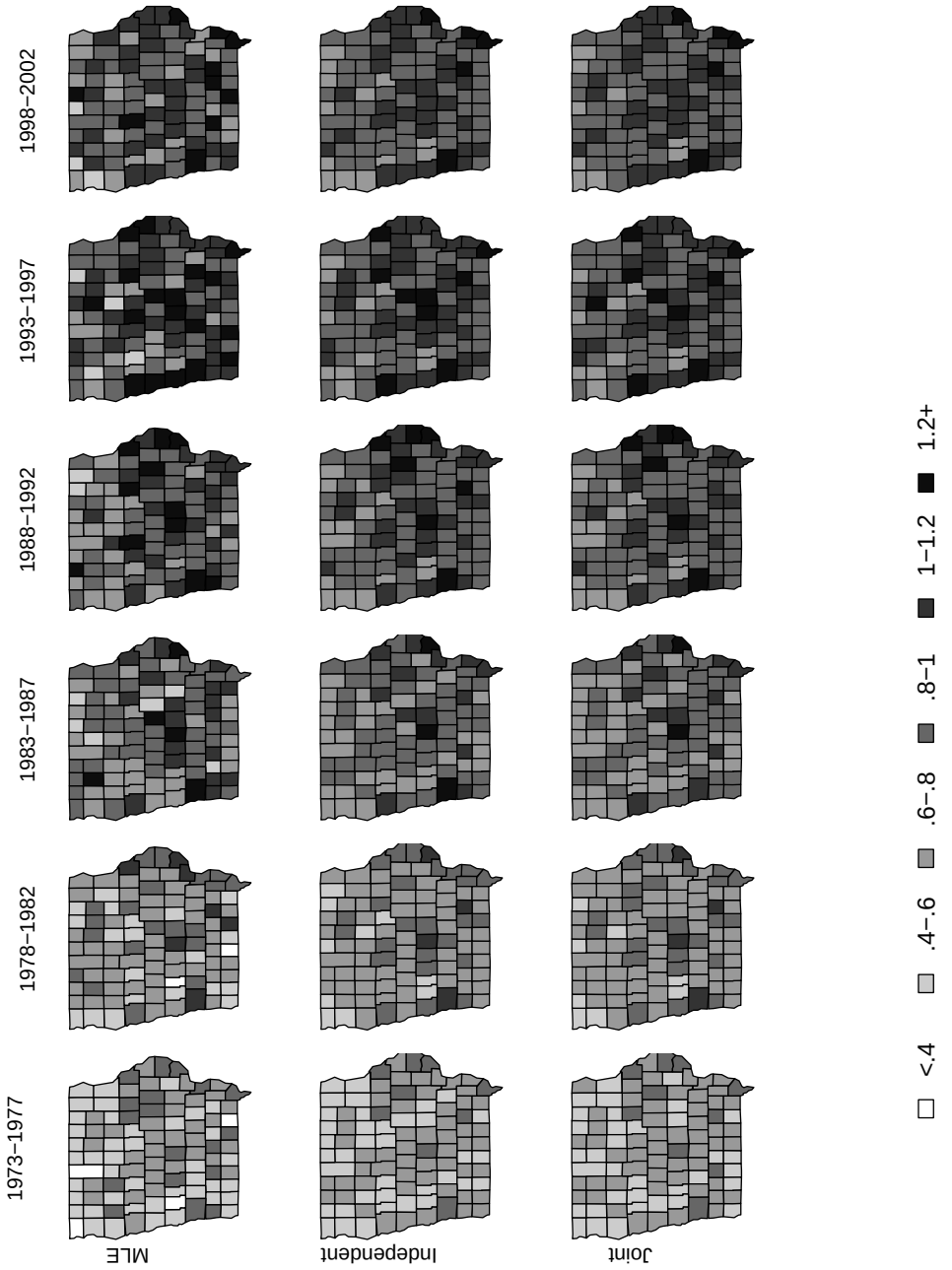

Figure 2: Lung cancer relative risks estimates by county and time period: Row 1 contains MLEs, Row 2 contains posterior means from the independent Bayesian model, and Row 3 contains posterior means from the joint model.

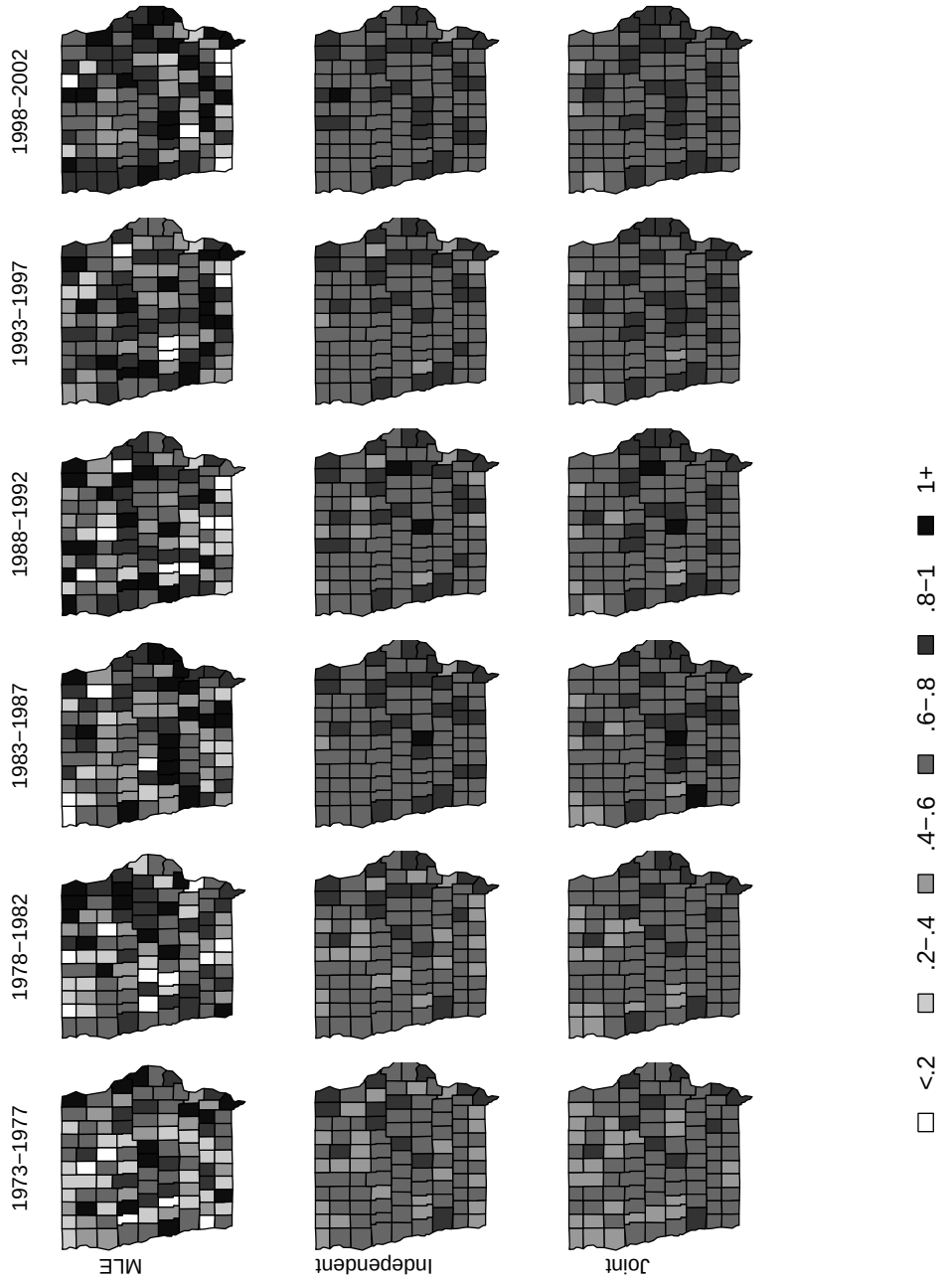

Figure 3: Oral cancer relative risks estimates by county and time period: Row 1 contains MLEs, Row 2 contains posterior means from the independent Bayesian model, and Row 3 contains posterior means from the joint model.

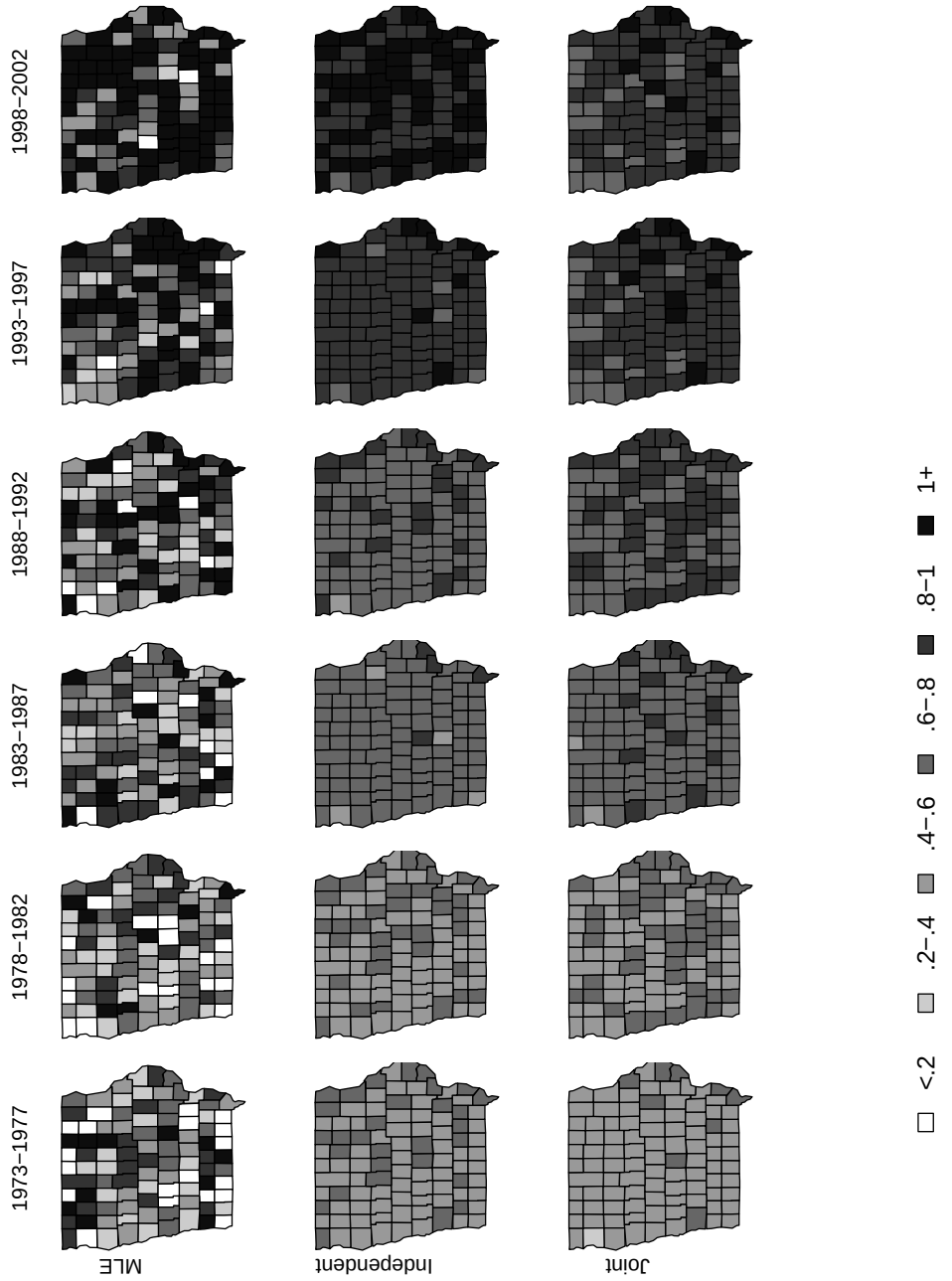

Figure 4: Esophageal cancer relative risks estimates by county and time period: Row 1 contains MLEs, Row 2 contains posterior means from the independent Bayesian model, and Row 3 contains posterior means from the joint model.

α_k is a disease specific intercept. The Z_i is a random spatial effect for county i whereas ϕ_k is a spatial scaling parameter for disease k. These are discussed further in Section 3.2. The t_j represents a random temporal effect of the j^{th} time period, and ψ_k is a temporal scaling parameter, as described in Section 3.3. Finally, the e_{ijk} are distributed as $e_{ijk} \overset{iid}{\sim} Normal(0, \delta_e)$ and allow for extra-Poisson variation due to risk factors not accounted for by the model.

3.2 Spatial correlation

Located in the first rows of Figures 2, 3, and 4 are the maximum likelihood estimates (MLEs) of the relative risks for lung, oral, and esophageal cancers, respectively. Spatial patterns are not readily apparent in these plots due to variability caused by small counts in the individual counties and time periods. We can make the assumption, though, that similarities exist in the spatial patterns because of risk factors shared among the three cancers. In the model, cancers are linked to a common correlation structure through the Z_i parameter. However, the ϕ_k scaling parameter allows the spatial variances to differ. Thus, we will be interested in examining if $\phi_1 = \phi_2 = \phi_3 = 1$. This equality would suggest no difference between the spatial structures of the three cancers. For identifiability, we set $\phi_1 = 1$ since lung cancer has the most county-level data.

The random spatial effects for counties follow the conditional autoregressive model (CAR) (Besag, 1974). Define a 99×99 adjacency matrix as $\mathbf{C} = (C_{uv})$ where $C_{uv} = 1$ if two counties u and v share a common boundary and 0 otherwise, with $C_{uu} = 0$. The vector of county effects is denoted as $\mathbf{Z} = (Z_1, \ldots, Z_{99})'$ and a 99×99 diagonal weighting matrix $\mathbf{D} = diag(d_i)$ with $d_i = \sum_j C_{ij}$. Then, the conditional distribution of Z_i is normally distributed with conditional mean and variance

$$E\big(Z_i|\mathbf{Z}_{-(i)}\big) = \frac{\rho_z}{d_i}\sum_{j=1}^{99} C_{ij}Z_j \qquad \text{and} \qquad Var\big(Z_i|\mathbf{Z}_{-(i)}\big) = \frac{\delta_z}{d_i},$$

respectively, where $\mathbf{Z}_{-(i)}$ denotes the vector of the spatial effects in all areas except the ith area. The parameter δ_z represents the degree of precision and controls the variability in Z_i while ρ_z symbolizes the degree of spatial association among regions. Sun *et al.* (2000) include the overall degree of spatial dependence parameter ρ_z in the CAR model. To ensure a proper (non-degenerate) distribution for $\mathbf{Z}$, ρ_z is bound between -1 and +1 by the theorem of diagonal dominant matrix (Graybill, 1983, p. 251).

3.3 Temporal correlation

Correlation over time can be seen by plotting the state-wide relative risk for each cancer since 1973 (see Figure 1(a)). There is a general increase from the 1973–1977 period to the 1998–2002 period in the relative risks for lung, oral, and esophageal cancer. Thus, a common latent temporal process represented as t_j is appropriate, although the variances of the temporal processes are allowed to vary across diseases through the model parameters ψ_k. We will examine if $\psi_1 = \psi_2 = \psi_3 = 1$; i.e., whether the cancers are homogenous with respect to their temporal structures.

We select an autoregressive model, AR(1), to model the temporal process across $j = 1, \ldots, 6$. The implementation is similar to that of Oleson and He (2004) and Kim and Oleson (2005), and we write

$$t_j = \rho_t t_{j-1} + \epsilon_j.$$

Assume that $\epsilon_j \sim Normal(0, \tau_j)$ is additional random noise with $\tau_1 = \delta_t/(1-\rho_t^2)$ and $\tau_j = \delta_t$ for $j = 2, \ldots, 6$.

This formulation assumes a stationary process. While the incidence rates have shown an increase since 1973, there is reason to believe that those rates will begin to decrease in the near future. This is evidenced by nationwide trends where the number of cigarettes sold in the U.S. is at the lowest number since 1961.

3.4 Bayesian model specification

A hierarchical Bayesian modeling approach provides a simplified conceptual framework to solve extremely complex problems involving covariance structures and latent spatial structures that would be particularly difficult to model otherwise. An additional advantage of the Bayesian paradigm is the ability to incorporate prior scientific knowledge in the analysis.

Priors were chosen to ensure a proper posterior distribution. We specified relatively vague $Normal(0, 100)$ priors for the mean parameter, α_k, and for the scaling parameters ϕ_k and ψ_k for $k = 2, 3$. $Gamma(.01, .01)$ priors for the inverse variance parameters are chosen. The correlation parameters ρ_z and ρ_t have $Uniform(-1, +1)$ priors.

In a fully Bayesian hierarchical model such as this, computation via numerical integration is not feasible. Instead we use Markov chain Monte Carlo (MCMC) techniques, such as Gibbs sampling, to simulate draws from the full conditional distributions. Many of the full conditionals in our model are normal or inverse-gamma which are easily sampled. Fortran 95 implementations of appropriate MCMC algorithms were developed and are available from the first author. Three

chains with dispersed starting values were generated from parallel runs of the MCMC sampler. Posterior inference was based on a total of 240,000 simulated draws, after discarding the first 20,000 iterations of each chain as a burn-in sequence. To monitor the convergence of our MCMC sampler, we used the diagnostics of Gelman and Rubin (1992) as well as graphical monitoring of the sample paths. All plots and maps were generated with R statistical software tools[2]. Convergence diagnostics and posterior summaries were performed with the BOA software[3]). The method of Chen and Shao (1999) was used to compute 95% highest posterior density (HPD) intervals.

4. Numerical Results

4.1 Maximum likelihood estimation

We begin by examining MLEs of θ_{ijk} for each of the cancers separately. These are simply the observed counts divided by the age-gender adjusted expected counts, i.e., $\hat{\theta}_{ijk} = Y_{ijk}/E_{ijk}$. The ratios are located in the first row of Figures 2–4 for lung, oral and esophageal, respectively. We have assumed similar spatial structures and temporal structures for these three cancers. Patterns are very difficult to discern from these maps, particularly for oral (Figure 3) and esophageal (Figure 4) cancers. Due to the small number of cases for these two cancers at the county level, individual estimates are not reliable thus making it difficult to recognize consistent spatial or temporal trends.

For example, we expect that neighboring counties will have similar relative risks, i.e. are spatially correlated. This type of pattern is not immediately recognizable when we examine the SMR MLEs for esophageal cancer (row 1 of Figure 4). In the map for 1973–1977, we see many instances where counties with the lowest category of relative risk (< 0.2) share a boundary with a county that had the highest category of relative risk (≥ 1.0). Eighteen of the nineteen counties found in the lowest relative risk category had zero cases of esophageal cancer in this time span. Many of the counties in the highest risk category are a product of a small number of incidences in sparsely populated counties rather than being a true high relative risk. This scenario is apparent when the neighboring counties have low risks. The same phenomenon is true for oral cancer as well.

The small counts also have a dramatic affect on temporal trends. Again, turn to esophageal cancer for an example. In the first row of Figure 4, we can see a general increase over time. Statewide, this general increase is true, but trends at the individual county-specific level are not as apparent. We illustrate

[2]See http://www.r-project.org.

[3]Smith, B. J. (2005). Bayesian Output Analysis Program (BOA), Version 1.1.5, http://www.public-health.uiowa.edu/boa.

with Mahaska County in southeast Iowa. Mahaska County is in the third row of counties from the bottom and the fourth from the right. The relative risk MLEs for the six time periods are 0.50, 0.50, 0.69, 0.00, 0.85, and 0.17. The corresponding counts during these time periods are 3, 3, 4, 0, 5, and 1 respectively. With the small changes in the number of incidences, the relative risks should not vary to the great extent that the MLEs do. Similar examples are present for oral cancer as well. These large fluctuations over space and time are not as drastic for lung cancer but similar disparities can certainly be seen.

4.2 Individual Bayesian modeling

Even with small counts and population sizes, a Bayesian model will be able to smooth the estimates for cancer relative risk. The results should show a more discernable spatial pattern and clearer temporal trends. To obtain independent Bayesian estimators for the cancer risks, equation (3.1) is used separately for $k = 1, 2$ and 3 with $\phi_k = 1$ and $\psi_k = 1$. The three models used are

$$\log(\theta_{ij(lung)}) = \alpha_{(lung)} + Z_{i(lung)} + t_{j(lung)} + e_{ij(lung)} \quad (4.1)$$

$$\log(\theta_{ij(oral)}) = \alpha_{(oral)} + Z_{i(oral)} + t_{j(oral)} + e_{ij(oral)} \quad (4.2)$$

$$\log(\theta_{ij(esophageal)}) = \alpha_{(esophageal)} + Z_{i(esophageal)} + t_{j(esophageal)} + e_{ij(esophageal)}. \quad (4.3)$$

Posterior summaries of the model parameters are provided in Table 1.

Table 1: Individual disease modeling: Bayesian posterior parameter estimates based on 240,000 MCMC samples.

	Lung Cancer		**Oral Cancer**		**Esophageal Cancer**	
Parameter	Mean	95% HPD	Mean	95% HPD	Mean	95% HPD
α	-0.198	(-0.572, 0.100)	-0.358	(-0.502,-0.223)	-0.275	(-0.707, 0.145)
δ_y	0.004	(0.002, 0.006)	0.007	(0.002, 0.014)	0.010	(0.002, 0.019)
δ_z	0.129	(0.088, 0.175)	0.165	(0.089, 0.247)	0.102	(0.019, 0.195)
δ_t	0.031	(0.004, 0.084)	0.011	(0.001, 0.028)	0.049	(0.005, 0.134)
ρ_z	0.595	(0.162, 0.985)	-0.527	(-0.990, 0.222)	-0.206	(-0.990, 0.715)
ρ_t	0.634	(-0.021, 0.999)	0.239	(-0.651, 0.989)	0.679	(0.087, 0.999)

The relative risk estimates from the independent models of (4.1) through (4.3) are mapped in the second row of Figures 2–4 for lung, oral and esophageal, respectively. The spatio-temporal modeling smoothed the risk estimates so that there is less county-to-county variability and patterns are more clearly seen. The

example from Section 4.1 of highly variable clustering has been removed. Neighboring counties now have similar relative risk estimates. In particular, there are no counties that fall into either the lowest category or the highest category in the first time period. Mahaska County shows a steadier trend across time as well with relative risk estimates of 0.51, 0.52, 0.61, 0.64, 0.80, and 0.85 for the six time periods. Note that there appears to be a higher pocket of relative risks for all three cancers near Pottawattamie County in western Iowa (third row of counties from the bottom in the far left) and surrounding Polk County in central Iowa (fourth row of counties from the bottom and the sixth from the left). Higher levels also appear in eastern Iowa. Lower levels of relative risk are seen in most of northwest Iowa for all three cancers.

Table 2: Joint Disease modeling: Bayesian posterior parameter estimates based on 240,000 MCMC samples.

Parameter	Mean	95% HPD
$\alpha_{(lung)}$	-0.182	(-0.443, 0.050)
$\alpha_{(oral)}$	-0.350	(-0.446,-0.269)
$\alpha_{(esophagus)}$	-0.345	(-0.714,-0.046)
$\delta_{y(lung)}$	0.005	(0.003, 0.007)
$\delta_{y(oral)}$	0.010	(0.003, 0.019)
$\delta_{y(esophagus)}$	0.011	(0.002, 0.023)
δ_z	0.130	(0.089, 0.174)
δ_t	0.028	(0.003, 0.075)
ρ_z	0.463	(-0.074, 0.930)
ρ_t	0.622	(-0.033, 0.999)
$\phi_{(oral)}$	0.869	(0.694, 1.051)
$\phi_{(esophagus)}$	0.704	(0.473, 0.942)
$\psi_{(oral)}$	0.280	(0.099, 0.462)
$\psi_{(esophagus)}$	1.323	(1.038, 1.611)

From Figure 1(a) we determined that relative risks had the same general increasing trend with a leveling off at the last time period. The time components, t_j, are plotted in Figure 1(b). The modeled patterns match the actual trends for the MLEs displayed in Figure 1(a). In addition, they all show the same general increasing pattern. A joint modeling of all three cancers will capture this information while the scaling parameters will allow for subtle differences.

The three individual spatial parameters, $\mathbf{Z}$, for each cancer are shown in Figures 1(c)–1(e). We see a similar underlying structure for these three cancers. There are low values in the northwest quadrant and much of northern Iowa. A

high pocket can be seen in central and eastern Iowa. These three maps do indeed show similar spatial trends and further support the use of a common latent spatial process in a joint model.

While we have smoothed the estimates to locate "hot-spots" in Iowa, there is still a great deal of variability in the estimates for the two lower risk cancers. A joint modeling of the three cancers should reduce the variability.

4.3 Joint Bayesian modeling

Finally, a simultaneous analysis of the three cancers was performed with the joint hierarchical Bayesian model developed in Section 3. Posterior summaries of the model parameters are provided in Table 2.

Significant temporal correlation resulted. A posterior mean of 0.622 was observed for ρ_t with a 95% HPD covering (-0.033, 0.999). In Figure 1(b), the posterior mean obtained for t_j in the joint analysis is plotted across time. As observed in the separate analyses, the cancer risks have been increasing over time with a recent leveling off. The joint values behave very similarly to the values from the lung cancer model. Not only does lung cancer have the most data, but the value in t_j for lung cancer generally falls between that of oral and esophageal cancers.

The temporal relationship is further evidenced by the estimates of ψ_k. A value of $\psi_k = 1$ implies that the temporal patterns are the same between the baseline disease and disease k. Lung cancer was used as the baseline because it had the most data to work with. The posterior mean for the esophageal scale parameter is 1.323, but the 95% HPD of $(1.038, 1.611)$ excludes unity. This suggests that it follows a similar temporal pattern to lung cancer but is somewhat more variable over time. Oral cancer exhibits a relatively attenuated trend that is characterized by a mean estimate of $\psi_{(oral)} = 0.280$ with a 95% HPD of $(0.099, 0.462)$. Overall, the results for the temporal scale parameters are consistent with the trend differences apparent in Figure 1(a). Returning to the figure, we note that esophageal cancer exhibits a slightly steeper increase than lung cancer, whereas oral exhibits a relatively flatter increase. The posterior estimates from the joint model indicate that although they have all shown increases, there are significant differences between the increases over time for the three cancers.

There are important associations over space as well. When $\phi_k = 1$, the spatial patterns are the same between the baseline disease and disease k. Recall that lung cancer is used as the baseline in our analysis. The mean for the esophageal scale parameter is 0.704 with a 95% HPD of $(0.473, 0.942)$. Oral cancer has an estimate of 0.869 $(0.694, 1.051)$ for $\phi_{(oral)}$. Unlike the case of temporal variance, oral and lung cancer appear to have very similar spatial variances.

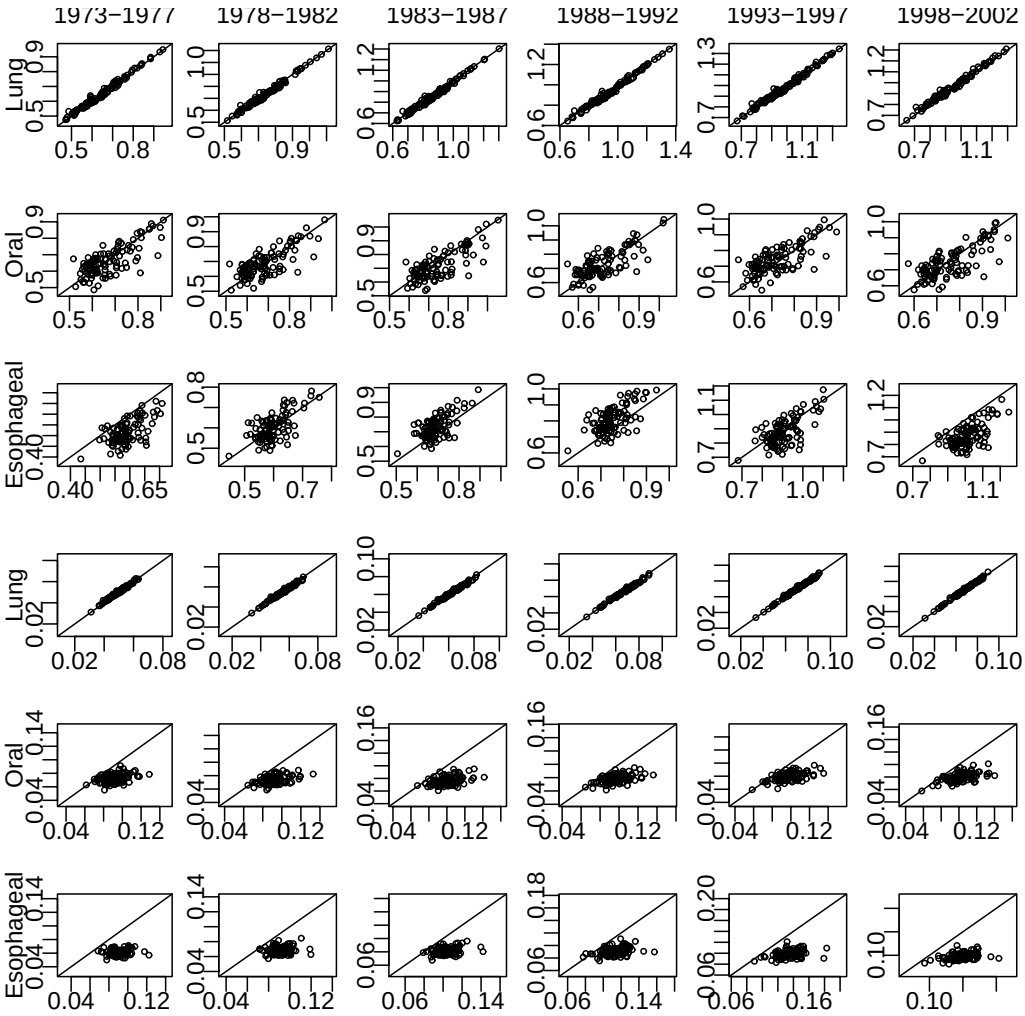

Figure 5: Comparison of relative risk posterior means (first three rows) and standard deviations (last three rows) between independent (horizontal axis) and joint Bayesian modeling (vertical axis) of lung, oral, and esophageal cancer by county and time period.

Next, consider the the risk estimates in the third row of Figures 2–4. There are many similarities between the separate modeling estimates and the joint modeling estimates. The lung cancer results are particularly similar. This is expected as there were a relatively large number of lung cancer cases in Iowa. The differences between the two methods will be found in examining the two lower risk cancers. Many counties had zero cases of esophageal or oral cancers during certain time periods. Each of the independent models smooth using information from neighboring counties and time periods for a single cancer. The joint model uses more knowledge by utilizing additional counts and spatio-temporal information for the two other cancers. We can now clearly see higher relative risks in south and southwest Iowa and lower risks in northwest Iowa.

Relative risk estimates from the individual modeling (horizontal axis) are plotted against the estimates from the joint modeling (vertical axis) in the first three rows of Figure 5. Lung cancer estimates found in the first row are relatively unaffected by the joint modeling. Oral cancer in the second row and esophageal cancer in the third row show larger county-level changes. The state averages appear to be similar, apart from the 1998–2002 period for esophageal cancer.

We expect the joint modeling to also reduce the variability in risk estimates. An examination of rows four through six of Figure 5 shows this to be the case. Similarly to the mean estimates, the variability in lung cancer estimates did not differ much between the separate and joint modeling approaches. The major increase in precision came for the lower risk diseases. We see dramatic reductions in variability for the oral and esophageal estimates in every time period.

5. Discussion

We have presented a model that incorporates joint information from multiple diseases sharing common disease risk factors. This model is beneficial when one or more diseases have small counts that lead to highly variable estimates. When a disease has small counts, the high variability in the small area estimates can make spatio-temporal patterns difficult to discern. Our approach has the potential to improve the precision of risk estimates through the joint modeling of related diseases. Information across different diseases is combined via the common spatial and temporal correlation structures specified in our model. Furthermore, we include scaling parameters to compare the strength of the spatial and temporal signals across diseases. The model is fit within a hierarchical Bayesian framework and so posterior inference can be performed on all model parameters and relative risks of interest. On the one hand, our model can improve upon the detection of "hot-spots" when different diseases are highly associated with the same risk factors. In such cases, the common spatial (temporal) structure could be used to identify regions that might benefit from risk reduction interventions. On the other

hand, the model can be used to determine how well a common spatial (temporal) structure describes patterns in different diseases. We found that there is a spatial relationship between lung, oral and esophageal cancers in Iowa. We also see a joint trend in time between the three diseases. Our proposed framework can easily incorporate any number of diseases while accounting for temporal correlation as well as spatio-temporal interaction.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B* **36**, 192-236.

Chen, M.-H. and Shao, Q-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* **8**, 69-92.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (Disc: 483-501, 503-511), *Statistical Science*, **7**, 457-472.

Graybill, F. A. (1983). *Matrices with Applications in Statistics*. Duxbury Press.

Held, L., Natário, I., Fenton, S. E., Rue, H. and Becker, N. (2005). Towards joint disease mapping, Statistical Methods in Medical Research **14**, 61-82.

Kim, H., Sun, D. and Tsutakawa, R. K. (2001). A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model, *Journal of the American Statistical Association* **96**, 1506-1521.

Kim, H. and Oleson, J. J. (2008). A Bayesian dynamic spatio-temporal interaction model: an application to prostate cancer incidence, *Geographical Analysis* **40**, 77-96.

Knorr-Held, L. and Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, Series A* **164**, 73-85.

Oleson, J. J. and He, C.Z., (2004). Space-time modeling for the Missouri Turkey Hunting Survey. *Environmental and Ecological Statistics* **11**, 85-101.

Oleson, J. J. and He, C. Z. (2007). Adjusting nonresponse bias at subdomain levels using multiple response phases. *Biometrical Journal* **49**, in press.

Sun, D., Tsutakawa, R. K. and Kim, H. and He, Z. (2000). Spatio-temporal interaction with disease mapping. *Statistics in Medicine*, **19**, 2015-2035.

Received September 6, 2006; accepted October 31, 2006.

Jacob J. Oleson
Department of Biostatistics
The University of Iowa
Iowa City, Iowa, USA
jacob-oleson@uiowa.edu

Brian J. Smith
Department of Biostatistics
The University of Iowa
Iowa City, Iowa, USA
brian-j-smith@uiowa.edu

Hoon Kim
Department of Mathematics & Statistics
California State Polytechnic University
Pomona, California, USA
hoonkim@csupomona.edu