An examination of disparities in cancer incidence in Texas using Bayesian random coefficient models

Corey Sparks

Disparities in cancer risk exist between ethnic groups in the United States. These disparities often result from differential access to healthcare, differences in socioeconomic status and differential exposure to carcinogens. This study uses cancer incidence data from the population based Texas Cancer Registry to investigate the disparities in digestive and respiratory cancers from 2000 to 2008. A Bayesian hierarchical regression approach is used. Specifically, a spatially varying coefficient model of the disparity between Hispanic and Non-Hispanic incidence is used. Results suggest that a spatio-temporal heterogeneity model best accounts for the observed Hispanic disparity in cancer risk. Overall, there is a significant disadvantage for the Hispanic population of Texas with respect to both of these cancers, and this disparity varies significantly over space. The greatest disparities between Hispanics and Non-Hispanics in digestive and respiratory cancers occur in eastern Texas, with patterns emerging as early as 2000 and continuing until 2008.

Title: An examination of disparities in cancer incidence in Texas using Bayesian random coefficient models.

Author: Corey S. Sparks[1,2]
[1]Department of Demography
The University of Texas at San Antonio
501 West Cesar E. Chavez Blvd
San Antonio, TX 78207
Email: corey.sparks@utsa.edu
Phone: 210 458 3166
Fax: 210 458 3164

[2]Department of Biostatistics and Epidemiology
The University of Texas Health Science Center at San Antonio
7703 Floyd Curl Drive
San Antonio, TX 78229, USA

Keywords: health disparity, Bayesian model, cancer incidence

Abstract: Disparities in cancer risk exist between ethnic groups in the United States. These disparities often result from differential access to healthcare, differences in socioeconomic status and differential exposure to carcinogens. This study uses cancer incidence data from the population based Texas Cancer Registry to investigate the disparities in digestive and respiratory cancers from 2000 to 2008. A Bayesian hierarchical regression approach is used. Specifically, a spatially varying coefficient model of the disparity between Hispanic and Non-Hispanic incidence is used. Results suggest that a spatio-temporal heterogeneity model best accounts for the observed Hispanic disparity in cancer risk. Overall, there is a significant disadvantage for the Hispanic population of Texas with respect to both of these cancers, and this disparity varies significantly over space. The greatest disparities between Hispanics and Non-Hispanics in digestive and respiratory cancers occur in eastern Texas, with patterns emerging as early as 2000 and continuing until 2008.

1. Introduction

Respiratory and digestive system cancers have been identified as often having direct and identifiable causal pathways associated with them, many of which are behavior or environmentally influenced. Lung cancer is perhaps the most widely recognized environmentally influenced cancer type, with strong evidence to support the effects of smoking, poor diet and direct inhalation of certain carcinogens, including asbestos and other indoor air pollutants (Alberg and Samet, 2003; Ruano-Ravina, Figueiras et al., 2003; Alberg, Ford et al., 2007). The exposure to these carcinogens generally leads to errors in somatic cell growth, such as chromosomal abnormalities, cellular mutations, and alterations in tumor suppressor cells. Gastrointestinal system cancers also have a variety of causes, with some consistency between the types of cancer, but other types also have distinct know etiologies. For example, hepatocellular carcinoma (primary liver cancer) has been directly linked with hepatitis infection, alcoholic cirrhosis and dietary aflotoxins (Stuver and Trichopoulous, 2008; El-Serag, 2012) while other digestive system cancers, such as colorectal cancers are heavily influenced by dietary and lifestyle factors (Chao, Thun et al., 2005). While the specific etiologies of the cancers of these two body systems sometimes have direct causal paths, they are generally thought to be influenced by both behavioral and environmental circumstances, which interact with familial and genetic pathways in complicated ways.

Disparities in cancer incidence and mortality exist between racial and ethnic groups in the United States and worldwide (Elmore, Nakano et al., 2005; Du, Fang et al., 2007; Vainshtein, 2008; Harper, Lynch et al., 2009; McKenzie, Ellison-Loschmann et al., 2010). The causes of these disparities have been suggested to be rooted in different

levels of socioeconomic status (SES), access to medical care, differential exposure to carcinogenic materials and differential treatment by medical staff of racial and ethnic minorities (Krieger, 2005; Sarfati, Blakely et al., 2006; Schootman, Lian et al., 2010). While these causes are often non-specific in their effects of how they influence cancer incidence, they do allow us to conceptualize and measure key factors related inequalities in health. Furthermore, understanding disparities in cancer risk and being able to visualize the place-based differences both in the determinants of cancer inequality can be a valuable tool to both scientist and policy maker alike.

The state of Texas is the second most populous state in the United States, with a current population estimate of 25.7 million persons. Between 2000 and 2010, Texas was the sixth fastest growing state, and the highest in total numerical population gain (Makun and Wilson, 2011). Additionally, it is consistently in the top five fastest growing states in the nation. The Hispanic population of Texas was estimated to be 9.1 million persons, or nearly 37% of the population in 2009 and Texas has the second largest Hispanic population, behind only California (Makun and Wilson, 2011). In addition to being a large part of the state's population, the Hispanic population also faces socioeconomic disadvantages compared to other ethnic groups. The poverty rate for Texas Hispanics was 25.8% according to the 2010 American Community Survey, while Non-Hispanic whites only had an 8.8% poverty rate (United States Department of Commerce, 2012).

For such a large and dynamic state, little population-based cancer disparity research has been published for Texas. In a recent study of cancer disparities in Texas counties, Phillips et. al. (2011) found that an index of socioeconomic well-being was significantly associated with county-level ratios of metastatic to non-metastatic tumors in

all-cause, female genital and lung cancers, although since a linear regression was used, no relative risks are available. In a study of El Paso county, Collins et. al. (2011) found higher cancer risk for the Hispanic population of that area, and they go on to discuss how in El Paso, areas of the city that had the highest levels of Hispanic population who had low levels of education had six times the risk of the more educated areas, and areas with the highest proportion of Hispanic renters had seven times the risk of cancer than other, more socioeconomically advantaged areas. Using a geographically weighted regression approach, (Tian, Wilson et al., 2011) on data from the Texas Cancer Registry, found not only that that Hispanics and Non-Hispanic Blacks faced disparities in breast cancer mortality, but that these disparities varied over space within the state. These studies likewise point to the placed-based inequality and increased risks that Hispanic populations face in certain areas within the state.

With respect to access-based disparities related to cancer risk Hispanics have been shown to have lower chances of seeking preventative care (Lantz, Mujahid et al., 2006; Shih, Zhao et al., 2006; Cristancho, Garces et al., 2008; Suther and Kiros, 2009; Hosain, Sanderson et al., 2011) of many different varieties including cancer screening. Reasons for not seeking care include lack of insurance, language barriers and the high cost of health care (Cristancho, Garces et al., 2008). In a study of colorectal cancer, Wan et. Al. (Wan, Zhan et al., 2012) found significant disparities for Hispanics and Non-Hispanic Blacks based accessibility to care.

1.2 Visualizing disparities across space

From a methodological standpoint, testing for disparities in rates is a relatively straightforward task and a variety of statistical procedures are well suited for it.

Specifically, a disparity in two rates can be measured as either a difference in total rates, or as a ratio of risks the groups being compared. In terms of visualizing the disparities, this can be more of a challenge. For measuring the disparity between population subgroups, the standardized risk ratio is a useful measure, but it is often subject to noise in the underlying rates, most notably in small populations or in cases of rare disease. Maps of such relative risks are, as a result of the noise caused by small populations, often lead to the reporting of unstable risk estimates. Tango (2010) describes a variety of methods for both visualizing and detecting disease clusters. Methods for mapping such risk ratios in a scan-statistic context have been described by Chen and co-authors (2008), and Bayesian disease mapping methods are also cited as being particularly good at mapping spatial disease risk (Lawson, Biggeri et al., 2000; Choo and Walker, 2008; Kim and Oleson, 2008; Lawson, 2009; Earnest, Beard et al., 2010).

It is the purpose of this paper to investigate the spatial variation in cancer incidence disparities between Hispanic and Non-Hispanic populations of the state of Texas between 2000 and 2009 using data derived from a population-based cancer registry. This research adds to the literature in spatial epidemiology by examining the disparities in these two populations over time and space by using a Bayesian modeling methodology, which models the variation in cancer disparities between these two populations within the state. The Bayesian modeling framework is used to specify a series of spatially varying coefficient models as a method of both more accurately modeling the disparity between these two populations, but also for visualizing where the disparities between the populations exist. The goal of this process it to provide a locally

accurate depiction of health disparities which state and local health officials could use in combating health inequalities.

2. Data and methods

2.1 Data source

Data for this analysis come from the Texas Cancer Registry's (www.dshs.state.tx.us/tcr/) Limited-Use data file from 2000 to 2008. Access to these data was approved by the Texas Department of State Health Services IRB #12-030. These data consist of de-identified individual records of primary cancer diagnoses by oncologists in the state of Texas. For the purposes of this study, relevant variables in the data include year of diagnosis, age, sex, Hispanic ethnicity, International Catalog of Disease for Oncology (ICD-O-3) codes for cancer diagnosis site and county of residence at the time of diagnosis. Two main types of cancer were chosen: digestive system (ICD-O-3 codes C150 – C488) and respiratory system cancers (codes C300 – C399). These cancers were chosen because several of the sub-types of these cancers have been linked to environmental or behavioral influences, and several have also been shown to vary between ethnic groups in their incidence (Wiggins, Becker et al., 1993; Singh and Siahpush, 2002; Howe, Wu et al., 2006; Singh and Hiatt, 2006; Willsie and Foreman, 2006). For the years of this study a total of n=155,652 digestive and n=124,438 respiratory system cases were in the data. The most prevalent form of digestive system cancer was colorectal cancer, with 53% of digestive cancers, and squamous cell carcinoma of the lung was the most prevalent respiratory cancer, representing 22% of all cases.

The dependent variable in the analysis is the count of either digestive or respiratory cancers in each of the 254 counties of Texas between 2000 and 2008. The data are stratified by ethnicity into two categories Hispanic and Non-Hispanic. The stratification of the cases is accomplished by using the Hispanic ethnicity variable in the registry. Thus for each year, there are two separate counts for each cancer type and for each of the 254 counties in the state. Since the dependent variables are counts, they are generally expressed as a standardized ratio of counts to expected counts. This is typically called the standardized incidence ratio (SIR), and is expressed:

$SIR_{ijk} = y_{ijk}/e_{ijk}$

Where y_{ijk} is the count of cases in the i^{th} county for the j^{th} year for the k^{th} ethnicity and e_{ijk} is the expected number of cases in the county for each group. Here, to estimate the expected number of cases for each county, year and ethnicity, an assumption of equal risks is used. To estimate the expected number of cases in each county, e_{ijk}, is calculated by assuming each county has the average incidence rate for the whole state for the period 2000 to 2008, or:

$e_{ijk} = \Sigma n_{ijk} * r_{ijk}$

where n_{ijk} is the number of residents in each county for each ethnicity, and r_{ijk} is the average incidence rate for the state for the period 2000 to 2008. This is repeated for each type of cancer: digestive and respiratory. This generates a set of expected values for the Hispanic and Non-Hispanic population of each county, using the statewide rate and the county population size for each group.

To control for background characteristics of the counties, and to measure proxies for factors affecting cancer risk, four independent variables are constructed. The first of

these is the metropolitan status of the county, which is measured as dummy variable indicating whether the county is considered metropolitan by the United States Department of Agriculture's Economic Research Service. These counties are coded as 1, and non-metro counties are coded as 0. The poverty rate in each county is calculated from the US Census Bureau's Summary File 3 for 2000, and is expressed as the proportion of all residents living below the poverty line in 1999. The proportion of the labor force in construction is used to measure a crude proxy for occupational exposure to certain carcinogens. This is again measured using the Census's Summary File 3 and expressed as a proportion. Finally, the Area Resource File (US Department of Health and Human Services, 2009) for 2008 is used to measure the number of hospitals in each county per 10,000 residents. This is used as a crude proxy for healthcare access in each county.

2.2 Statistical methods

2.2.1 Model Specification

Since the dependent variable is a count, the outcome is assumed to follow a Poisson distribution. To model this outcome, a log-linear Poisson hierarchical regression model for each county, i, year, j, ethnicity, k, and type of cancer, C, is specified as:

$$y_{Cijk} \mid \theta_{Cijk} \sim \text{Poisson}\ (e_{Cijk} * \theta_{Cijk})$$

The relative risk function, θ_{Cijk}, can be parameterized using a number of different models, the present paper considers a Bayesian model specification.

In the Bayesian modeling paradigm all model parameters are considered to be random variables and are given a prior distribution and all inference about these parameters is made from the posterior distribution of these parameters, given the

observed data and the information given in the priors. This is generally referred to as Bayes Theorem, and typically stated as:

$p(\theta|y) \propto p(y|\theta)p(\theta)$

Where $p(\theta|y)$ is the posterior distribution of the model parameter of interest, $p(y|\theta)$ is the model likelihood function, here defined as a Poisson likelihood, and $p(\theta)$ is the prior distribution for the parameter of interest. Inference for all parameters is done via their posterior distribution, which can be used to derive mean values, quantiles or other descriptive statistics. One useful method for summarizing these distributions is the Bayesian Credible Interval (BCI), not unlike frequentist confidence interval, which gives the values of the posterior density for each parameter that contain 100*(1-α)% of the posterior density. Inference on these BCI regions usually consists of examining if the null hypothesis value of the parameter, typically zero, is contained in the interval.

Since the primary interest in this paper is the relative difference between the incidence of cancer in the Hispanic and Non-Hispanic populations of each county, the simplest way to parameterize the model is as a linear difference in the incidence rates using a simple, unstructured linear predictor. This is the first model considered, and is parameterized and given priors as:

$$\begin{aligned}
&\ln \theta_{Cijk} = \alpha_C + \delta_C * eth_{Cik} + X'\beta_C + u_{Ci} + t_{Cj} \\
&\alpha_C \sim \text{flat} \\
&\delta_C \sim N(0,.0001) \\
&\beta_C \sim N(0,.0001) \\
&u_{Ci} \sim N(0, \tau_{uC}) \\
&t_{Cj} \sim N(t_{j-1}, \tau_{tC})
\end{aligned} \qquad \text{(Model 1)}$$

which models the relative risk as a linear function of a grand intercept for each cancer type, α_C, a mean difference between the two ethnicities for each cancer type, δ_C, a linear predictor effect of the independent variables for each cancer type, $X'\beta_C$, an unstructured heterogeneity term for each county and cancer type, u_{Ci}, which is equivalent to a random-intercept model, and a temporally correlated random effect for each year and cancer type, t_{Cj}. In this model there is a single parameter for measuring the disparity between Hispanics and non-Hispanics for each cancer type, and this is done on average for the entire state. This model additionally captures the underlying characteristics of the counties, and the temporal correlation between years in the relative risk. Priors are assigned to all parameters in a minimally informative fashion, with an improper flat prior for α_C, high variance Normal distribution priors for the δ_C and β_C and u_i, a random-walk Normal distribution prior for t_j and Uniform distributions for the standard deviations of the unstructured heterogeneity and temporal autocorrelation components.

A second model adds more flexibility to Model 1 by including a random slope for each county's difference between Hispanic and non-Hispanic risk. This model is specified as:

$$
\begin{aligned}
&\ln \theta_{Cijk} = \alpha_C + \delta_{Ci} * eth_{Cik} + X'\beta_C + u_{Ci} + t_{Cj} \\
&\alpha_C = flat \\
&\delta_{Ci} \sim N(0, \tau_{\delta C}) \\
&\beta_C \sim N(0, .0001) \\
&u_{Ci} \sim N(0, \tau_{uC}) \\
&t_{Cj} \sim N(t_{j-1}, \tau_{tC})
\end{aligned}
\qquad \text{Model 2}
$$

which is similar to (1), but includes a δ_{Ci} term which allows the differences between Hispanic and non-Hispanic risk to vary between counties, equivalent to a random-slopes model. This is much like the spatially varying coefficient model discussed elsewhere

(Gelfand, Kim et al., 2003; Banerjee, Carlin et al., 2004), but in this case the spatially varying coefficient measures the disparity between the two subpopulations. While this model itself is not new, the application of it to a health disparities outcome is new contribution. Again, priors are assigned to all parameters in a minimally informative fashion, with an improper flat prior for α_C, high variance Normal distribution priors for the δ_C and β_C and u_i, a random-walk Normal distribution prior for t_j and Uniform (0,100) distributions for the standard deviations of the random δ slope, unstructured heterogeneity and temporal autocorrelation components.

To adequately model the spatial and temporal structure of the data, a space-time interaction model is also specified (Model 3). This model was first described by Bernardinelli and coauthors (Bernardinelli, Clayton et al., 1995) with the model structure elaborated on by Knorr-Held (Knorr-Held, 2000) and used in a diverse array of settings by other authors (Ugarte, Goicoa et al., 2009; Lawson, Song et al., 2010; Schrodle and Held, 2011; Ugarte, Goicoa et al., 2012). In the present setting, two space-time interaction models are specified. The first uses an unstructured spatio-temporal interaction term and adds this to the structure in Model 2. This random interaction term follows the form of the Type 1 interaction discussed by Knorr-Held {Knorr-Held, 2000 #346}. The second spatio-temporal model uses the same unstructured spatio-temporal interaction term and a spatially correlated random effect term to allow for spatial autocorrelation in overall risk between neighboring counties. These models include terms for the spatio-temporal heterogeneity in general risk, as compared to Models 1 and 2 which focus more on the general disparity between Hispanics and Non-Hispanics. These models are referred to as Models 3 and 4 and have forms:

$$\ln \theta_{Cijk} = \alpha_C + \delta_{Ci} * eth_{Cik} + X' \beta_C + u_{Ci} + t_{Cj}$$
$$\alpha_C = flat$$
$$\delta_{Ci} \sim N(0, \tau_{\delta C})$$
$$\beta_C \sim N(0, .0001)$$
$$u_{Ci} \sim N(0, \tau_{uC})$$
$$t_{Cj} \sim N(t_{j-1}, \tau_{tC})$$
$$\psi_{Cij} \sim N(0, \tau_{\psi C})$$

Model 3

Where a space-time interaction effect is added to the linear predictor, and given a vague zero-mean Normal prior distribution, and a random slope to the disparity parameter is also included. The final model adds a spatially varying general intercept, in contrast to the unstructured intercepts in the other three models to account for spatial variation in total cancer risk:

$$\ln \theta_{Cijk} = \alpha_C + \delta_{Ci} * eth_{Cik} + X' \beta_C + u_{Ci} + t_{Cj}$$
$$\alpha_C = flat$$
$$\delta_{Ci} \sim N(0, \tau_{\delta C})$$
$$\beta_C \sim N(0, .0001)$$
$$u_{Ci} \sim N(\frac{1}{n} \sum_{j \sim i} u_{Cj}, \tau_{uC} / n_i)$$
$$t_{Cj} \sim N(t_{j-1}, \tau_{tC})$$
$$\psi_{Cij} \sim N(0, \tau_{\psi C})$$

Model 4

, which is again specified with the same space-time random effect as Model 3, but changes the u_{Ci} to a spatially structured conditionally autoregressive (CAR) random

effect. The notation for the CAR prior symbolizes that each county has a random effect which is the mean of its spatial neighbors, not including location *i*, which is symbolized by the term *j~i* in the equation.

For geographic modeling, neighbors are identified using a first order Queen contiguity rule. Other neighbor specifications were examined, specifically a first order rook contiguity rule, and the results were substantively robust to this other neighbor specification. Also, since the precision terms for Bayesian hierarchical models have been shown to be sensitive to prior specifications, a sensitivity analysis is performed. The models specified above all considered Uniform distributions for the standard deviation of each of the precision parameters. To examine the sensitivity of the models to alternative specifications, proper Gamma (.5, .0005) priors are also considered for all precision terms. This prior distribution has been used by other authors, and is thought of to be a sufficiently vague prior for the precision for these parameters.

2.3 Computing

The software R (R Development Core Team, 2010) and the R package R2OpenBUGS (Sturtz, Ligges et al., 2005) were used to prepare data for analysis and OpenBUGS 3.2.1 was used for parameter estimation. To estimate the posterior densities of model parameters, samples were drawn from their stationary posterior distribution via multiple chain Markov-Chain Monte Carlo simulation in OpenBUGS. A total of 20,000 iterations were generated, and the first 10,000 were discarded as a burn in period for the Markov chains. Results were derived from another 100,000 samples from the stationary distribution of the parameters, using every 10^{th} sample to avoid autocorrelation in the Markov Chains. Two Markov chains were started at divergent ends of the parameter

space and convergence was assessed using the Gelman-Rubin diagnostic (Gelman and Rubin, 1992), which showed convergence by 10,000 iterations. This generated a total of 10,000 samples for each parameter in the model from each chain. Combining the output from both chains yielded 20,000 samples for each stochastic node in the model. To summarize the posterior distributions of the model parameters, posterior means and 95% credible intervals are calculated. Five models were examined as specified in 2.2.1. Model fit and improvement is assessed between the models with the Deviance Information Criterion (DIC) (Spiegelhalter, Best et al., 2002). The DIC measures the penalized deviance of each model, with the penalty term representing the model's estimated number of parameters. DIC is typically calculated as $DIC = \bar{D} + pD$, where $\bar{D} = -2\sum_{1}^{G} l(y \mid \theta^{G}) / G$ is the average model deviance evaluated after the Markov Chains have converged, and *pD* is an approximate number of parameters in the model. Since the DIC is a measure of overall model fit and complexity, but not a real measure of model performance, the mean absolute prediction error or MAPE for each model is also computed. This is calculated:

$$MAPE = \frac{1}{n}\sum\left|y_i - y_i^{\text{Pred}}\right|$$

, where y_i^{Pred} is a simulated value from the posterior predictive distribution of cancer cases, based upon posterior Poisson means for each model. MAPE serves as a general measure of the predictive capacity of the models, where the model with the lowest total error, or *MAPE* would be the preferred model. It is not, however, penalized for over parameterization, so it is feasible for a model with a higher DIC to have a lower MAPE.

3. Results

3.1 Descriptive Results

Descriptive statistics for the dependent variables and the predictors are presented in Table 1.

[TABLE 1 HERE]

A gradual increase in the average number of cases per county is observed over the nine years in the analysis. Also, we see that many more cases of both types of cancer (on average) occur to Non-Hispanics than to Hispanics. It should be noted that for digestive cancers, between 25 (2005) and 36% (2000) of counties had a zero count for Hispanic digestive cancer cases and between 38 (2003) and 46% *(2002) had a zero count for Hispanic respiratory cancer cases. Also presented in Table 1 are the observed average risk ratios for the state for each year. These are calculated as ratio of the observed SIR for Hispanics (SIR_H) and the observed SIR for Non-Hispanics (SIR_{NH}) for each year. For digestive cancers, every year shows a qualitative elevated risk for Hispanics compared to Non-Hispanics, and all years except 2000 show an elevated risk of respiratory cancer for Hispanics. Likewise, respiratory cancers show a consistent trend of higher risk in Hispanics, but not to the same risk level as digestive cancers. With respect to the predictor variables, in 2000 nearly 18 percent of the population of Texas was in poverty, with a wide degree of variation as seen by the inter quartile range. On average there were .66 hospitals per 10,000 people in each county in the state, and there were sixty-five counties with no hospitals. Slightly over 8 percent of the work force was employed in construction, and the USDA considered thirty percent of counties in the state to be metropolitan.

3.2 Results of Bayesian models

Table 2 presents the posterior means of the independent variables in the four models described above. Also, 95% Bayesian credible intervals are provided for each parameter. Model fit statistics are also provided at the bottom of the table for each model. Lastly, the model variance components are provided.

[Table 2 HERE]

Across the four models, some of the fixed predictors show similar patterns. For digestive cancers, the poverty rate shows a negative association with Hispanic relative risk in Models 1 through 3, and only Model 1 shows a significant effect of hospitals per capita, which shows a negative association with the relative cancer risk. This suggests that in areas of higher poverty, the average disparity is lower. Also, when the simpler model specification is considered, accessibility to medical care decreases (as measured only by hospitals per capita) the digestive cancer disparity. Model 3 also shows a significant association between the disparity and the proportion of the workforce in construction, which is shown to increase the Hispanic relative risk. Interestingly, Model 4 shows no significant effects of the predictor variables. This suggests that the predictors used may just be proxies for the spatial concentration in risk measured by the spatial random effect in Model 4.

Respiratory cancer incidence is affected consistently by two of the predictors. The proportion of the work force in construction is positively associated with respiratory cancer risk in three of the models, potentially suggesting an occupation-specific risk pattern. Similar to the digestive cancer outcome, the county poverty rate shows a significant negative association with Hispanic risk in three of the four models. This

suggests that without controlling for spatially correlated heterogeneity in risk, the poverty rate is protective in terms of the Hispanic disparity, or that Hispanics and Non-Hispanics in high poverty areas face similar levels of risk.

When the four models are compared using the standard DIC, Model 4 shows the best model fit with a DIC 1,240 points lower than Model 1, 910 points lower than Model 2 and 650 DIC points lower than Model 3. Strong evidence is present that Model 1 is not adequate to describe the patterns of Hispanic/Non-Hispanic disparities in either cancer, as every other model shows large drops in DIC, and noticeable drops in MAPE between the models. When comparing Models 2 and 3, strong evidence also exists for adding the spatio-temporal random effect to Model 3, again with a large drop in DIC and reduction in MAPE. The comparison between Model 3 and 4 again shows strong evidence, in terms of the DIC, for the use of the spatially structured model intercepts, but the MAPE shows little to no difference from Model 3. It should be noted that the model deviance in Model 4 is actually higher than Model 3, but the pD estimate shows Model 4 being a much more parsimonious model, hence generating a lower DIC.

Turning to the Hispanic disparity parameters, for Model 1, the Hispanic/Non-Hispanic disparity was measured as a fixed effect in the model. For both cancer types, results show a significant increase in relative risk for the Hispanic population relative to the Non-Hispanic population, on average across the state. For the other models, the coefficients of the modes are best presented graphically, as each county has an estimate for the disparity for each cancer type. These estimates are presented in Figure 1 as posterior mean estimates of the Hispanic disparity in relative risk ($e^{\delta C}$) for each county for models 2 through 4.

[Figure 1 Here]

Figure 1 displays the estimates of the Hispanic disparity in relative risk parameters for each of the three varying coefficient models (Model 2-4). The first row of Figure 1 shows the parameters for respiratory cancers, while the second row of the figure shows the estimates for digestive cancers. Model 2, which was the independent, unstructured disparity parameter model shows isolated counties with high excess Hispanic risk throughout the state, but a general West to East gradient, with higher relative risks (above 1) in the eastern and southern portions of the state. Model 3 is the central column of Figure 1 and, much like Model 2 also shows elevated risk throughout the state, with perhaps come clusters of counties in the central and eastern portions of the state. The third column of Figure 1 illustrates the disparity parameter from Model 4, and this effect shows less structure to the disparity parameter than either Models 2 or 3, most likely resulting from the addition of the spatially correlated intercept term in the model.

The value of these figures is that the actual disparity in risk is being visualized, which shows us where within the state public health officials might try to focus activities in order to reduce the disparity in risk between these two populations.

3.3 Spatio-temporal Relative Risk Estimation

Figure 2 displays the estimated Hispanic relative risk for digestive cancers (e^{θ}) generated from Model 4.

[Figure 2 Here]

The quantity being mapped is the linear predictor of the Poisson mean, with all random effects included, which is interpreted as the model-based standardized incidence ratio (SIR). Each panel in the figure shows the spatial pattern for one year of relative risk

between 2000 and 2008. We see a general concentration of elevated Hispanic digestive cancer risk in the eastern portion of the state, as evidenced by relative risks greater than one (darker blue in color). This pattern is consistent, if not increasing over time, with more counties showing greater Hispanic relative risk over time. Lower risk for Hispanics occurs in North and Western Texas, and also along the border with Mexico, except for a few counties in extreme South Texas in the latter time periods.

[Figure 3 Here]

Figure 3 provides the complementary space-time risk map for the respiratory cancer outcome. Again, we see higher Hispanic risk in Eastern Texas, but perhaps a more concentrated pattern, compared to the digestive cancer maps. Also present is the lower risk in North and West Texas, as seen in Figure 2 for digestive cancers. Figure 3 also highlights a consistent spatial cluster of high risk in extreme East Texas for a cluster of three to five counties located North of Harris county (city of Houston). These counties include Montgomery, Liberty, San Jacinto, Walker, Polk and Orange. These counties are quite rural and have low proportions of Hispanic residents (average of 9.3%, or about 8,900 Hispanic persons on average per county).

Finally, the sensitivity analysis of alternative priors for the model random effects showed very close agreement between the vague Uniform and the vague Gamma (.5, .0005) prior distributions. Since Model 4 showed evidence of being the best fitting model, the sensitivity analysis focused on its estimates. The precision point estimates for the temporal random effects (τ_t) for the digestive and respiratory cancers, respectively were 1,581 and 2,117 from the Gamma prior and 1,715.0 and 2,691.0 from the Uniform. Likewise, the precisions for the correlated heterogeneity (τ_u) were 53.6 and 18.9 for the

Gamma prior and 57.3 and 24.2 for the Uniform. The precisions for the varying disparity parameter were 22.4 and 25.3 from the Gamma and 142.3 and 125.5 from the Uniform prior. The precisions for the spatio-temporal random effect (τ_{ψ}) were 177.0 and 269.5 for the Gamma prior model and 292.0 and 285.5 for the Uniform prior model. While this is only one model, the overlap between the precisions is strong enough to validate the results. The one notable difference is the random effect for the disparity between the two populations, which showed a lower precision in the Gamma prior model.

4. Discussion

This paper illustrated the application of Bayesian varying coefficient models to the study of cancer incidence disparities between the Hispanic and Non-Hispanic population of Texas over the period 2000 to 2008. This paper adds to the literature in health disparities by using advanced Bayesian statistical methods to investigate the spatial non-stationarity of health disparities in two major form of cancer incidence. The primary goal of the analysis was to examine the usefulness of the spatially varying coefficient model (Gelfand, Kim et al., 2003; Banerjee, Carlin et al., 2004) within the Bayesian modeling framework using a variety of model specifications, including models that included interactions between space and time. Alternative model specifications structured the disparity in incidence between the two subpopulations differently, from a fixed effect on the grand mean to a spatially varying coefficient model for each county in the state. The flexibility of the Bayesian framework also allowed for the models to be compared using standard model complexity criteria (DIC) and a measure of predictive loss (MAPE).

The model that best fit the data was the space-time model with a spatially varying intercept and an unstructured random slope for the disparity between Hispanics and Non-Hispanics, according to the minimum DIC criteria. Some differentiation between models was found using the MAPE, but as seen in Model 4, MAPE and DIC did disagree to very small extent, probably because the MAPE calculation does not penalize over-parameterized models. This suggests that the disparity between Hispanics and Non-Hispanics in these two cancer types is best modeled through a count-specific approach, and an approach which allows for general spatially structured variation in risk. This also suggests that there are counties within the state where the Hispanic population is at higher risk for both of these cancers, but these counties do not necessarily occur close to one another spatially. There are notable exceptions to this, however, with the a small group of rural counties in Eastern Texas showing a strong spatio-temporal cluster of Hispanic risk for the respiratory cancer outcome.

Overall, a general disparity in terms of both cancers for Hispanics was found, where they face higher risk for both digestive and respiratory cancers than the Non-Hispanic population of the state. Significant effects were found on total cancer risk consistently including the county poverty level, and the proportion of the workforce in construction. The labor force composition finding makes sense, as workers in construction industries often face higher levels of exposure to airborne particulates that could increase cancer risk. The finding for the county poverty rate was that in areas with higher poverty, the overall relative risk of cancer was lower, and deserves more discussion. This effect was seen for both cancer types, in all but the final model (Model 4), and is in stark contrast to findings from national data (Singh, Miller et al., 2003) for

many types of cancer, which show higher incidence and mortality in both Hispanics and Non-Hispanics in areas with higher poverty. Two important considerations need to be made concerning the differences reported by Singh et. al. and the present study. First, Singh et. al. did not use data from Texas, and the time period for the present study is later than those considered in their report. It is possible that the experience of the Texas population is different from the national average; such local variations are common in health research. Secondly, since this model was formulated in terms of the relative disadvantage for Hispanics compared with Non-Hispanics, the poverty effect may make perfect sense. If the effect is really interpreted as the effect of poverty on the difference in relative risk between these populations, then we can say that poverty is acting as an equalizer, where the differences between the two ethnic groups is smaller in areas of higher poverty. This would agree well with the findings from Singh et. al.'s piece, and again can be indicative that populations, on average, living in poverty face overall higher risk, but the disparity between populations is less. Also, the effect was not found in the final model (Model 4), which included a spatially varying intercept, which affects the total cancer risk. Thus the poverty effect may, in the non-spatial models, have been indicative of some general pattern of differential risk that was geographically structured, and modeled directly in Model 4 using the spatially correlated term.

Further research is needed to investigate the specifics of the counties identified in the analysis as having excess Hispanic cancer risk. This can be done by a more localized analysis of the individual-level data this analysis is derived, and by investigating housing conditions, access to healthcare and potential environmental contaminants in these areas directly. Such ecological analyses as that presented here are rarely truly informative for

individual cancer diagnoses, but they can be very influential in terms of public health activities to reduce cancer disparities at the population level.

References

Alberg AJ, JG Ford, et al. Epidemiology of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007; **132**: 29S-55S.

Alberg AJ and JM Samet Epidemiology of lung cancer. Chest 2003; **123**: 21S-49S.

Banerjee S, BP Carlin, et al. Hierarchical modeling and analysis for spatial data. Boca Raton: CRC/ Chapman & Hall; 2004.

Bernardinelli L, D Clayton, et al. Bayesian analysis of space-time variation in disease risk. Stat Med 1995; **14**: 2433-2443.

Chao A, MJ Thun, et al. Meat consumption and risk of colorectal cancer. JAMA 2005; **293**: 172-182.

Chen J, RE Roth, et al. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of US cervical cancer mortality. Int J Health Geogr 2008; **7**:

Choo L and SG Walker A new approach to investigating spatial variations of disease. J Roy Stat Soc a Sta 2008; **171**: 395-405.

Collins TW, SE Grineski, et al. Understanding environmental health inequalities through comparative intracategorical analysis: Racial/ethnic disparities in cancer risks from air toxics in El Paso County, Texas. Health Place 2011; **17**: 335-344.

Cristancho S, DM Garces, et al. Listening to rural Hispanic immigrants in the midwest: A community-based participatory assessment of major barriers to health care access and use. Qual Health Res 2008; **18**: 633-646.

Du XL, S Fang, et al. Racial disparities and socioeconomic status in association with survival in a large population-based cohort of elderly patients with colon cancer. Cancer-Am Cancer Soc 2007; **110**: 660-669.

Earnest A, JR Beard, et al. Small area estimation of sparse disease counts using shared component models-application to birth defect registry data in New South Wales, Australia. Health Place 2010; **16**: 684-693.

El-Serag HB Epidemiology of Viral Hepatitis and Hepatocellular Carcinoma. Gastroenterology 2012; **142**: 1264-1273.

Elmore JG, CY Nakano, et al. Racial inequities in the timing of breast cancer detection, diagnosis, and initiation of treatment. Med Care 2005; **43**: 141-148.

Gelfand AE, H Kim, et al. Spatial modeling with spatially varying coefficient processes. J Am Stat Assoc 2003; **98**: 387-396.

Gelman A and D Rubin Inference from iterative simulation using multipel sequences (with discussion). Statistical Science 1992; **7**: 457-511.

Harper S, J Lynch, et al. Trends in Area-Socioeconomic and Race-Ethnic Disparities in Breast Cancer incidence, Stage at Diagnosis, Screening, Mortality, and Survival among Women Ages 50 Years and Over (1987-2005). Cancer Epidem Biomar 2009; **18**: 121-131.

Hosain GMM, M Sanderson, et al. Racial/Ethnic Differences in Predictors of Psa Screening in a Tri-Ethnic Population. Cent Eur J Publ Heal 2011; **19**: 30-34.

Howe HL, XC Wu, et al. Annual report to the nation on the status of cancer, 1975-2003, featuring cancer among US Hispanic/Latino populations. Cancer-Am Cancer Soc 2006; **107**: 1711-1742.

Kim H and JJ Oleson A Bayesian dynamic spatio-temporal interaction model: An application to prostate cancer incidence. Geogr Anal 2008; **40**: 77-96.

Knorr-Held L Bayesian modelling of inseparable space-time variation in disease risk. Stat Med 2000; **19**: 2555-2567.

Krieger N Defining and investigating social disparities in cancer: critical issues? Cancer Cause Control 2005; **16**: 5-14.

Lantz PM, M Mujahid, et al. The influence of race, ethnicity, and individual socioeconomic factors on breast cancer stage at diagnosis. Am J Public Health 2006; **96**: 2173-2178.

Lawson AB Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Boca Raton: CRC Press; 2009.

Lawson AB, AB Biggeri, et al. Disease mapping models: an empirical evaluation. Stat Med 2000; **19**: 2217-2241.

Lawson AB, HR Song, et al. Space-time latent component modeling of geo-referenced health data. Stat Med 2010; **29**: 2012-2027.

Makun P and S Wilson (2011). Population Distribution and Change: 2000 to 2010. Washington D.C., U.S. Department of Commerce.

McKenzie F, L Ellison-Loschmann, et al. Investigating reasons for socioeconomic inequalities in breast cancer survival in New Zealand. Cancer Epidemiol 2010; **34**: 702-708.

Philips BU, G Gong, et al. Correlation of the ratio of metastatic to non-metastatic cancer cases with the degree of socioeconomic deprivation among Texas counties. Int J Health Geogr 2011; **10**:

R Development Core Team (2010). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Ruano-Ravina A, A Figueiras, et al. Lung cancer and related risk factors: an update of the literature. Public Health 2003; **117**: 149-156.

Sarfati D, T Blakely, et al. Patterns of disparity: ethnic and socio-economic trends in breast cancer mortality in New Zealand. Cancer Cause Control 2006; **17**: 671-678.

Schootman M, M Lian, et al. Temporal trends in area socioeconomic disparities in breast-cancer incidence and mortality, 1988-2005. Breast Cancer Res Tr 2010; **122**: 533-543.

Schrodle B and L Held Spatio-temporal disease mapping using INLA. Environmetrics 2011; **22**: 725-734.

Shih YCT, LR Zhao, et al. Does Medicare coverage of colonoscopy reduce racial/ethnic disparities in cancer screening among the elderly? Health Affair 2006; **25**: 1153-1162.

Singh GK and RA Hiatt Trends and disparities in socioeconomic and behavioural characteristics, life expectancy, and cause-specific mortality of native-born and

foreign-born populations in the United States, 1979-2003. Int J Epidemiol 2006; **35**: 903-919.

Singh GK, BA Miller, et al. (2003). Area Socioeconomic Variations in U.S. Cancer Incidence, Mortality, Stage, Treatment, and Survival, 1975–1999. NCI Cancer Surveillance Monograph Series. Bethesda, MD, National Cancer Institute. **4**.

Singh GK and M Siahpush Ethnic-immigrant differentials in health behaviors, morbidity, and cause-specific mortality in the United States: An analysis of two national data bases. Hum Biol 2002; **74**: 83-109.

Spiegelhalter DJ, N Best, et al. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society Series B-Methodological 2002; **64**: 583-639.

Sturtz S, U Ligges, et al. R2WinBUGS: A Package for Running WinBUGS from R. Journal of Statistical Software 2005; **12**: 1-16.

Stuver S and D Trichopoulous (2008). Cancer of the liver and biliary tract. Textbook fo Cancer Epidemiology. Adami H, D Hunter and D Trichopoulous. Oxford, Oxford University Press**:** 308-332.

Suther S and GE Kiros Barriers to the use of genetic testing: A study of racial and ethnic disparities. Genet Med 2009; **11**: 655-662.

Tango T Statistical Methods for Disease Clustering. New York: Springer; 2010.

Tian N, JG Wilson, et al. Spatial association of racial/ethnic disparities between late-stage diagnosis and mortality for female breast cancer: where to intervene? Int J Health Geogr 2011; **10**: 24.

Ugarte MD, T Goicoa, et al. Testing for space-time interaction in conditional autoregressive models. Environmetrics 2012; **23**: 3-11.

Ugarte MD, T Goicoa, et al. Evaluating the performance of spatio-temporal Bayesian models in disease mapping. Environmetrics 2009; **20**: 647-665.

United States Department of Commerce (2012). American Factfinder 2.

US Department of Health and Human Services (2009). Area Resource File (ARF) 2008-2009. Health Resources and Services Administration Bureau of Health Professions. Rockville, MD.

Vainshtein J Disparities in breast cancer incidence across racial/ethnic strata and socioeconomic status: A systematic review. J Natl Med Assoc 2008; **100**: 833-839.

Wan N, FB Zhan, et al. Access to healthcare and disparities in colorectal cancer survival in Texas. Health Place 2012; **18**: 321-329.

Wiggins CL, TM Becker, et al. Cancer Mortality among New-Mexico Hispanics, American-Indians, and Non-Hispanic Whites, 1958-1987. J Natl Cancer I 1993; **85**: 1670-1678.

Willsie SK and MG Foreman Disparities in lung cancer: Focus on Asian Americans and Pacific Islanders, American Indians and Alaska Natives, and Hispanics and Latinos. Clin Chest Med 2006; **27**: 441-+.

Table 1. Descriptive statistics for dependent and independent variables used in the analysis.

Cancer Type and Year	Mean # Cases	IQR	Mean # Cases (Non-Hispanic)	Mean # Cases (Hispanic)	Mean SIR_H/SIR_{NH}
Digestive Cancer Cases per County					
2000	30.9	18	49.9	12.0	0.87
2001	32.2	18	51.8	12.6	1.44
2002	32.9	19	52.6	13.2	1.18
2003	33.7	19.25	53.5	14.0	1.14
2004	34.4	22	54.0	14.8	1.31
2005	34.8	22	53.9	15.8	1.32
2006	35.2	21	54.3	16.1	1.30
2007	36.1	23	55.8	16.4	1.46
2008	36.1	20	55.1	17.0	2.06
	155,652 total cases				
Respiratory Cancer Cases per County					
2000	25.6	15	46.0	5.2	1.28
2001	26.5	17	47.2	5.8	1.42
2002	26.9	17	48.2	5.6	1.16
2003	27.8	17	49.4	6.1	1.62
2004	27.6	16.25	49.2	5.9	1.18
2005	28.1	17	49.9	6.4	1.48
2006	27.4	16	48.4	6.5	1.67
2007	27.8	16	48.7	6.8	1.61
2008	27.2	15	48.1	6.4	1.54
	123,438 total cases				
Predictors	**Mean**	**IQR**			
% in Poverty	17.76	6.58			
Hospitals/10,000 People	0.66	0.79			
% in Construction	8.11	3.15			
% Metro Counties	30.31	1.00			

n=254 counties

Table 1(on next page)

Table 2 Results for the alternative Bayesian model specification parameters.

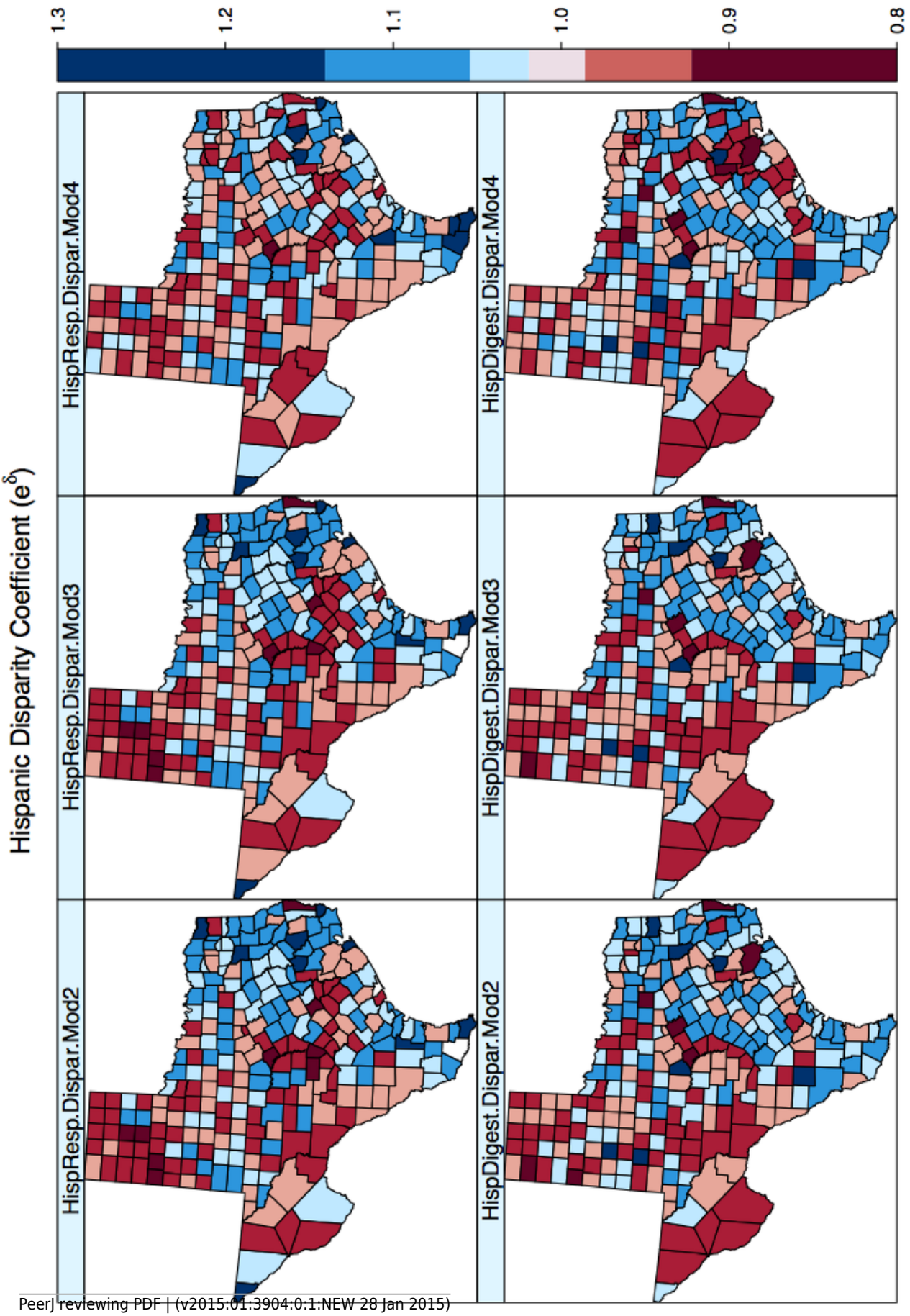
Table 2. Results for the alternative Bayesian model specification parameters.

	Model 1		Model 2		Model 3		Model 4	
Parameter	95% Credible Interval		95% Credible Interval		95% Credible Interval		95% Credible Interval	
α	**-0.212 (-0.274 , -0.147)**	**-0.213 (-0.277 , -0.151)**	**-0.181 (-0.248 , -0.114)**	**-0.122 (-0.188 , -0.061)**	**-0.185 (-0.257 , -0.120)**	**-0.115 (-0.178 , -0.057)**	**-0.191 (-0.263 , -0.124)**	**-0.131 (-0.195 , -0.073)**
Predictors, *β*	Digestive	Respiratory	Digestive	Respiratory	Digestive	Respiratory	Digestive	Respiratory
% in Poverty	**-0.043 (-0.061 , -0.026)**	**-0.027 (-0.051 , -0.002)**	**-0.044 (-0.067 , -0.019)**	**-0.028 (-0.054 , -0.002)**	**-0.044 (-0.066 , -0.022)**	**-0.041 (-0.069 , -0.011)**	-0.009 (-0.038 , 0.020)	0.012 (-0.023 , 0.049)
Hospitals per capita	**-0.036 (-0.061 , -0.011)**	-0.018 (-0.048 , 0.013)	-0.026 (-0.055 , 0.004)	-0.015 (-0.047, 0.019)	-0.026 (-0.056 , 0.002)	-0.010 (-0.045 , 0.024)	-0.011 (-0.040 , 0.019)	-0.007 (-0.404 , 0.025)
% in Construction	0.009 (-0.007 , 0.013)	**0.066 (0.043 , 0.089)**	0.021 (-0.001 , 0.043)	**0.067 (0.043, 0.091)**	**0.020 (0.000 , 0.042)**	**0.076 (0.050 , 0.104)**	0.006 (-0.017 , 0.028)	**0.063 (0.036 , 0.090)**
Metro County	0.026 (-0.011 , 0.062)	0.040 (-0.011 , 0.091)	0.004 (-0.001 , 0.049)	0.036 (-0.014, 0.087)	0.004 (-0.039 , 0.049)	0.005 (-0.052 , 0.061)	0.010 (-0.031 , 0.053)	0.039 (-0.011 , 0.090)
Hispanic Disparity	**0.050 (0.036 , 0.064)**	**0.103 (0.083 , 0.123)**	See Figure 1		See Figure 2		See Figure 3	
Model Fit								
Deviance ($\bar{D}$)	40,620		39,948		39,469		39,490	
DIC	40,950		40,461		40,360		39,710	
pD	337.4		512.3		886.4		225.7	
MAPE	8.776		8.501		8.025		8.027	
Variance Components								
τ_t	1752.7	2486.6	1733.0	2470.0	1725.7	2774.5	1715.0	2691.0
τ_u	113.7	48.0	124.5	55.7	131.6	57.4	57.35	24.2
τ_δ	-	-	127.4	93.6	126.9	93.5	142.3	125.5
τ_ψ	-	-	-	-	292.6	281.2	292.0	285.5

*Parameters in bold type represent estimates whose credible intervals do not contain 0.

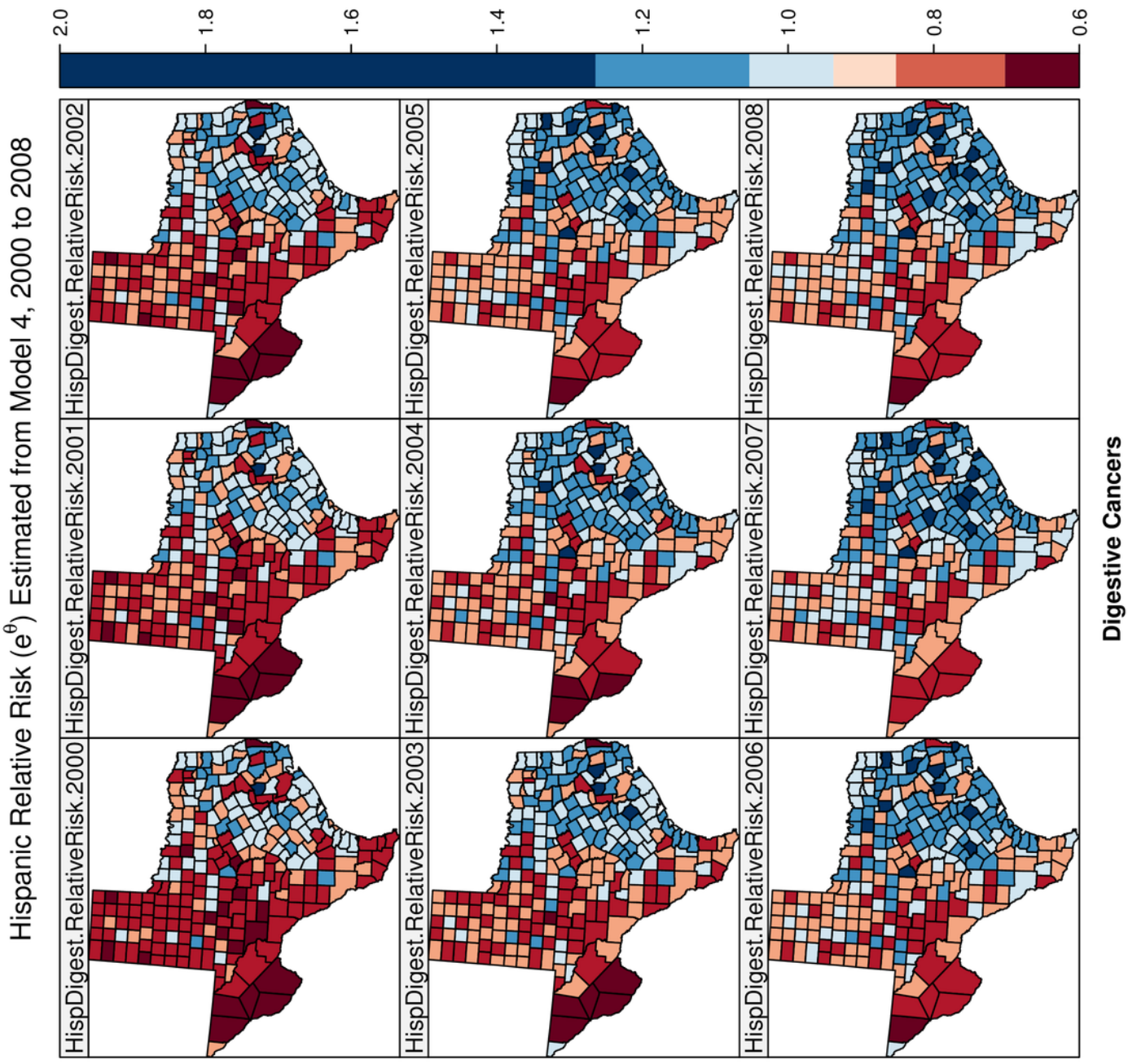
1

. Map showing the distribution of Hispanic/Non-Hispanic disparity parameter (δ) for Models 2-4.


Hispanic Disparity Coefficient (e^{δ})
1.3
1.2
1.1
1.0
0.9
0.8
HispResp.Dispar.Mod4
HispDigest.Dispar.Mod4
HispResp.Dispar.Mod3
HispDigest.Dispar.Mod3
HispResp.Dispar.Mod2
HispDigest.Dispar.Mod2

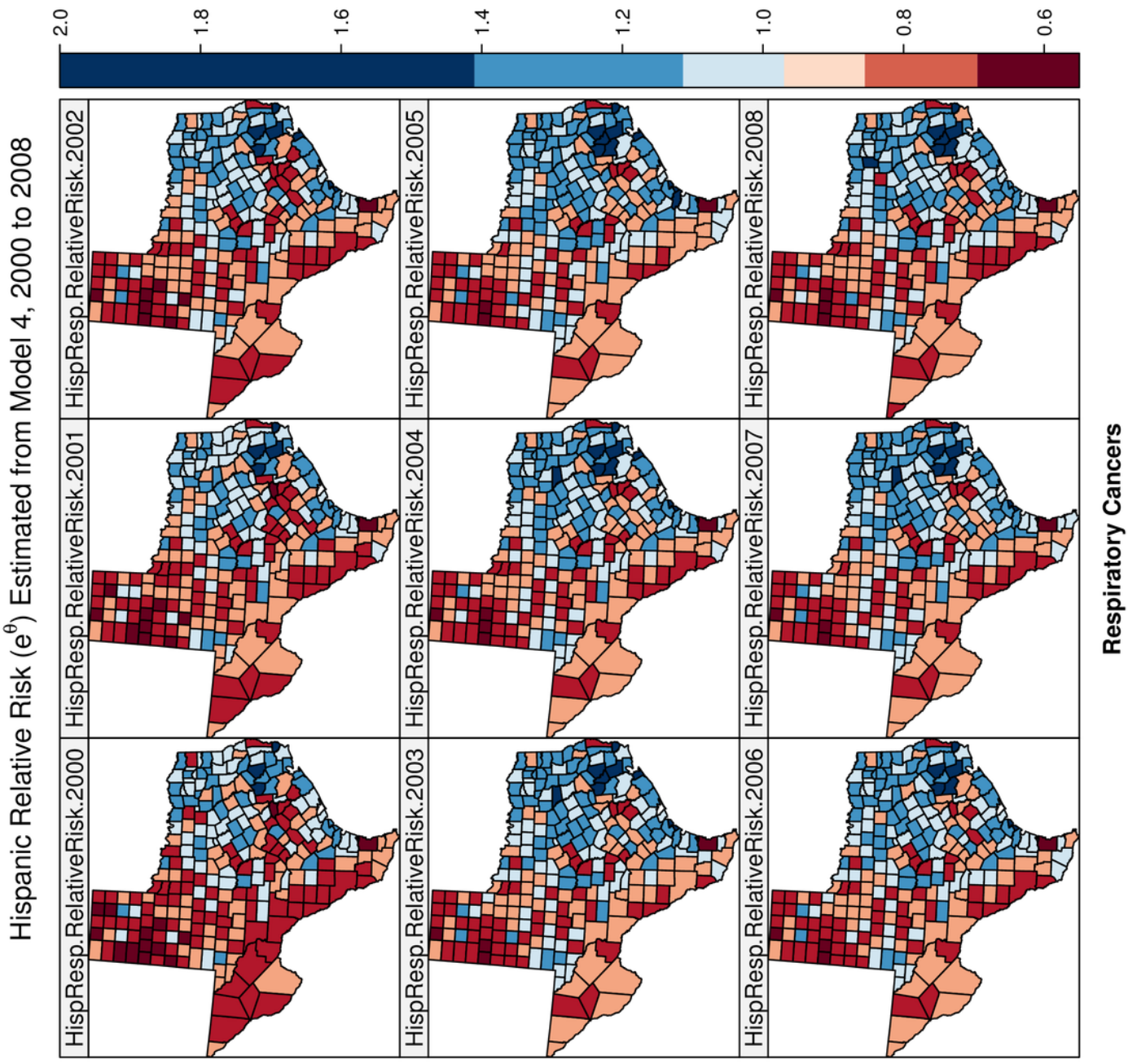
2

Maps showing the distribution of Hispanic relative risk for digestive cancers, derived from Model 4, 2000- 2008.


Hispanic Relative Risk (e^{θ}) Estimated from Model 4, 2000 to 2008
HispDigest.RelativeRisk.2000
HispDigest.RelativeRisk.2001
HispDigest.RelativeRisk.2002
HispDigest.RelativeRisk.2003
HispDigest.RelativeRisk.2004
HispDigest.RelativeRisk.2005
HispDigest.RelativeRisk.2006
HispDigest.RelativeRisk.2007
HispDigest.RelativeRisk.2008
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
Digestive Cancers

3

Maps showing the distribution of Hispanic relative risk for respiratory cancers, derived from Model 4, 2000- 2008.


Hispanic Relative Risk (e^{θ}) Estimated from Model 4, 2000 to 2008
HispResp.RelativeRisk.2000
HispResp.RelativeRisk.2001
HispResp.RelativeRisk.2002
HispResp.RelativeRisk.2003
HispResp.RelativeRisk.2004
HispResp.RelativeRisk.2005
HispResp.RelativeRisk.2006
HispResp.RelativeRisk.2007
HispResp.RelativeRisk.2008
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
Respiratory Cancers