Published in final edited form as:

Stat Med. 2008 August 15; 27(18): 3612–3628. doi:10.1002/sim.3232.

Joint spatial survival modeling for the age at diagnosis and the vital outcome of prostate cancer

Huafeng Zhou¹, Andrew B. Lawson¹, James R. Hebert^{1,2,3}, Elizabeth H. Slate², and Elizabeth G. Hill²

¹Department of Epidemiology and Biostatistics, the Arnold School of Public Health, University of South Carolina, 800 Sumter Street, Columbia, SC 29208, U.S.A

²Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, 135 Cannon Street, Suite 303, Charleston, SC 29425, U.S.A

³Cancer Prevention and Control Program, University of South Carolina, 2221 Devine Street, Columbia, SC 29208, U.S.A

Summary

Prostate cancer is the most common malignancy in men and a leading cause of cancer mortality among males in the United States. Large geographical variation and racial disparities exist in both the incidence of prostate cancer and in the survival rate after diagnosis. In this population-based study, a joint spatial survival model is constructed to investigate factors that affect the age at diagnosis of prostate cancer and the subsequent survival. The joint model for these two time-to-event outcomes is specified through parametric models for age at diagnosis and survival time conditional on diagnosis age. To account for possible correlation in these outcomes among men from the same geographical region, frailty terms are included in the survival model. Both spatially correlated and uncorrelated frailties are incorporated in each model considered. The deviance information criterion (DIC) is used to select a best-fitting model within the Bayesian framework. The results from our final best-fitting model indicate that race, marital status at diagnosis, and cancer stage are significantly associated with both of the two time-to-event outcomes. No pattern emerged in the geographical distribution of age at prostate cancer diagnosis. In contrast, a spatially clustered pattern was observed in the geographic distribution of survival experience post diagnosis.

Keywords

Prostate cancer; joint spatial survival model; spatial clustering; deviance information criterion (DIC)

1. Introduction

In public health studies, data often are collected by geographic region or location of the medical services provider. Because geographically closer regions usually share similar environmental and social factors, data from the same or closer regions tend to be more alike than those from more distant regions. Each region or location where individual observations arise is often treated as a cluster, and cluster-specific random effects are introduced to account for this spatial correlation.

For the analysis of clustered survival data, the Cox proportional hazards model has been extended to a Cox frailty model by the introduction of random, cluster-specific effects acting additively on the log hazard function [1, 2]. Often these random effects are assumed to be independent and identically distributed (*i.i.d.*) normal random variables with zero mean. In some situations, however, the validity of this *i.i.d.* assumption may be questionable, e.g., neighboring or geographically closer regions may share a common feature and hence the cluster-specific random effects may exhibit spatial dependence. Consequently, Cox frailty models that impose a spatial correlation structure on the random effects have been developed recently [3–6]. The cluster-specific random effect captures the influence of an unmeasured risk factor common to all individuals in the same region. In the case of *i.i.d.* random effects, this risk factor displays no spatial pattern; spatially correlated random effects reflect a risk factor that is more similar for geographically closer regions. In practice, the choice between an *i.i.d.* random effect and a spatially correlated random effect is often unclear. The inclusion of both types of random effects provides a prior distribution intermediate between independence and a purely local spatially structured dependence of the random effects [7].

Multiple event times are observed on a single subject in many survival studies. An example is cohort studies in which the date of diagnosis of disease and, subsequently, the date of death caused by that disease are both recorded. In this situation, there are two time-to-event outcomes: the time from birth to disease diagnosis and the time from disease diagnosis to death. Usually, the latter time is dependent on the former, e.g., age is an important risk factor for many diseases and the survival after disease diagnosis is dependent on age.

Interest often lies in the association of key covariates with both of the event-specific survival times. A simple analysis of this type of data fits survival models for the two time-to-event outcomes separately and, to reflect their dependence, uses time to diagnosis as a covariate in the model for post-diagnosis survival. The effect of covariates on each of the event-time outcomes is then investigated and reported from the two models independently. Another natural choice, and one that better accommodates the bivariate nature of the data, is a joint model for the two time-to-event outcomes simultaneously. However, this option is not common in practice, even though joint modeling is now widely used in the setting of analyzing survival data with repeated measures in longitudinal studies [2, 8] and in spatial epidemiology for joint disease mapping [9]. One obvious advantage of joint modeling over separate modeling is that the joint modeling will allow simultaneous investigation of the effect of covariates on the two related time-to-event outcomes and will be more efficient.

Prostate cancer (PrCA) is the most common malignancy in men and a leading cause of cancer mortality among males in the United States [10]. Large geographical variation and racial disparities exist in both the incidence of prostate cancer and in survival after diagnosis [11–19]. With the growing recognition of the merit of investigating spatial variation in prostate cancer incidence, mortality, treatment, and understanding survival patterns to inform cancer control and prevention, the application of Geographic Information System (GIS) techniques to prostate cancer research attracts increasing attention. This is reflected by the recent special issue of the *American Journal of Preventive Medicine* [20] on prostate cancer and GIS. The state of South Carolina (SC) has one of the highest prostate cancer mortality rates in the world, currently ranking 3rd in the nation. Although this high mortality rate might be partly attributable to the high proportion of African Americans in SC, it is notable that the mortality rate of prostate cancer among African American men in SC is markedly higher than that among African American men in the rest of the US as a whole [21–22]. The incidence rate of prostate cancer has a tendency to cluster spatially in SC. There are many counties in the eastern and southern part of the state where incidence rates are significantly higher than the state average [21–22].

Our main objective in this study is to characterize the spatial distributional pattern of the risk of prostate cancer diagnosis and survival experience after diagnosis. We also assess the effects of race, marital status at diagnosis and cancer stage at diagnosis on both age at diagnosis and survival following diagnosis.

2. The data

Results presented here are based on the analysis of prostate cancer data obtained from the South Carolina Central Cancer Registry (SCCCR) for the period 1997–2001. The SCCCR is a population-based cancer incidence registry covering the entire state of South Carolina (SC) that consistently receives the highest (gold) rating from the North American Association of Central Cancer Registries, indicating data completeness in excess of 97.5%. In addition to incidence data, mortality data for prostate cancer patients were obtained for this study from the Division of Vital Records of the South Carolina Department of Health and Environment Control.

Our raw data set contains all incident cases of prostate cancer diagnosed during the years 1997–2001 in SC. The individual-specific information for a patient that are used in this study are: age (age of the patient at diagnosis in complete years), race (patient's race as recorded in medical records: White, Black, or Other, with Black being the designator for African American), county (patient's county of residence at the time of diagnosis), stage (Surveillance, Epidemiology, and End Results (SEER) summary stage 2000 at the initial diagnosis[23]), date of diagnosis, marital status at diagnosis, vital status (vital status of the patient as of 12/31/03), and date of death (available if death happened before 12/31/2003, otherwise censored). The age at diagnosis and the survival time after diagnosis are the two event-time outcomes to be modeled in this study. The survival time after diagnosis is calculated as the time period, in months, between date of diagnosis and date of death (if censored, then 12/31/2003 is the censoring time).

There were 14,607 prostate cancer cases diagnosed in the 5-year period from 1997–2001 in SC. Observations with missing values on race, subject's county of residence, or marital status at diagnosis were excluded in this analysis. In our raw data set, 86 cases with a race subgroup of "other" also were excluded in this study. After the exclusion of observations with missing values on the above 3 covariates, there were 45 individuals for whom the date of diagnosis was coincident with the date of death. These 45 observations also were excluded from the analysis. This left a total of 12,238 observations in the final analytic data set.

Table 1 provides a summary of the characteristics of the prostate cancer patients included in this study. African-American males comprise just under 1/3 of all diagnosed cases during the 5-year study period of 1997–2001. To provide context for this figure, note also that almost all (99.9%) of our diagnoses are among men aged ≥ 45 years, and that among *all* White and Black men aged ≥ 45 years in SC, only 22.5% were Black during these 5 years [24], reflecting their younger age distribution. Thus, in SC, African-American males have a markedly higher age-adjusted incidence rate of prostate cancer than their European (i.e., White) counterparts.

In our SC data, approximately 5% of the diagnoses were late-stage (distant) disease and 6% were unstaged; the corresponding distribution from the SEER 9 areas* was 5% for distant stage and 4% for unstaged during the period 1995–2001 [25]. In the SEER report, localized and regional stage were combined, thus the data for these two stages of cancer is not

---

*The SEER 9 areas are San Francisco, Connecticut, Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, and Atlanta.

differentiable. In our data, the median age at diagnosis is 68 years for White males and 67 years for African-American males, while during the similar period of time, the corresponding numbers for the total US are 69 years for White males and 66 years for African-American males. In our study, death was observed among 18.38% of all prostate cancer patients before the study end date (12/31/2003). Unfortunately, we lack the medical evidence to conclude that the primary cause of death is prostate cancer. We should bear this in mind when interpreting the results in this study. Thus, it is appropriate to say “death after a diagnosis of PrCA,” not “death due to PrCA”.

A referee has noted that the 5-year window 1997–2001 of PrCA cases included in our study induces a truncation on the time-to-diagnosis outcome. In particular, left truncation exists because a man's PrCA diagnosis is not included in our study unless he survived without PrCA until 1997. Thus, our observed time to diagnosis for a man born in year b is truncated below by $1997 - b$. This truncation would create bias in conclusions regarding time to PrCA onset in the full SC population, as is particularly evident if one considers the extreme case where all diagnoses included arise from men in the same birth cohort. Of course, our data contain a mixture of birth cohorts (range 1897–1962) so that the observed onset time distribution is a corresponding mixture of truncated distributions. Nonetheless, the results of our study should be interpreted only in the context of PrCA diagnoses occurring in 1997–2001.

3. Method

3.1. Model specification

We used a common parametric survival distribution, the Weibull, which is very flexible, though simple, and is widely used for modeling survival data in biomedical science and reliability data in engineering. It is the only model that has both proportional hazard and accelerated failure time properties. Here, we built our model based on the proportional hazard property of the Weibull distribution.

Let T_{1ij} denote the age at diagnosis and T_{2ij} denote the survival time after diagnosis for patient j in county i , and let t_{1ij} and t_{2ij} respectively, denote the realized values of T_{1ij} and T_{2ij} . The joint modeling of T_{1ij} and T_{2ij} is realized as the product of the conditional distribution of T_{2ij} given T_{1ij} and the marginal distribution of T_{1ij} :

$$f_{T_{1ij}T_{2ij}}(t_{1ij}, t_{2ij}) = f_{T_{1ij}}(t_{1ij}) \cdot f_{T_{2ij}|T_{1ij}}(t_{2ij}|t_{1ij}).$$

Our interest is inference concerning parameters in both $f_{T_{1ij}}$ and $f_{T_{2ij}|T_{1ij}}$. We assume a Weibull distributional form for $f_{T_{1ij}}$ and $f_{T_{2ij}|T_{1ij}}$ that is

$$f_{T_{1ij}} \sim \text{Weibull}(\gamma_1, \lambda_{1ij}) \quad \text{and} \quad f_{T_{2ij}|T_{1ij}} \sim \text{Weibull}(\gamma_2, \lambda_{2ij}|t_{1ij}).$$

The Weibull distribution is completely determined by two parameters: the shape γ and the scale λ . Although the covariate effects could be modeled through the mean of the Weibull distribution, which is a function of the scale and shape parameters, or through either of the parameters, modeling via the scale parameter is most often used. The covariate effects are modeled through the scale parameters as given by the two equations below,

$$\ln(\lambda_{1ij}) = \alpha_1 + \beta_1^T x_{1ij} + W_{1i} \quad \text{and} \\ \ln(\lambda_{2ij}) = \alpha_2 + \beta_2^T x_{2ij} + \beta_{2t} * t_{1ij} + W_{2i},$$

where x_{1ij} , x_{2ij} respectively, denote the individual-specific covariates (such as race, marital status, cancer stage) included in each of the two event-time models. Note that, for the event-time T_{2ij} model, t_{1ij} , the realized value of T_{1ij} , is included as a covariate to reflect the dependence of the survival experience after diagnosis on age at diagnosis. W_{1i} and W_{2i} denote the county-specific random effects.

In contrast to the classical Cox proportional hazard models, a parametric form for the baseline hazard function is assumed here. This proportional hazard frailty model accommodates intracluster (i.e. within-county) correlation when making inference on the individual-specific covariates. Note that conditionally on the frailty terms, the event times $\{(T_{1ij}, T_{2ij})\}$ within a cluster are independent. Based on the specifications of the survival model for T_{1ij} and T_{2ij} , the corresponding hazard functions are respectively,

$$h_1(t_{1ij}; x_{1ij}) = \gamma_1 t_{1ij}^{\gamma_1 - 1} \exp(\alpha_1 + \beta_1^T x_{1ij} + W_{1i}) \quad \text{and} \\ h_2(t_{2ij} | t_{1ij}; x_{2ij}) = \gamma_2 t_{2ij}^{\gamma_2 - 1} \exp(\alpha_2 + \beta_2^T x_{2ij} + \beta_{2t} * t_{1ij} + W_{2i}),$$

and the joint likelihood function of the data is given by

$$L(\alpha, \beta, W, \gamma_1, \gamma_2; \mathbf{t}, \mathbf{x}) \propto \prod_{i=1}^I \prod_{j=1}^{n_i} \{ [\lambda_{1ij} \gamma_1 t_{1ij}^{\gamma_1 - 1} \exp(-\lambda_{1ij} t_{1ij}^{\gamma_1})] \times [\lambda_{2ij} \gamma_2 t_{2ij}^{\gamma_2 - 1} \exp(-\lambda_{2ij} t_{2ij}^{\gamma_2})]^{\delta_{ij}} \times [\exp(-\lambda_{2ij} t_{2ij}^{\gamma_2})]^{1 - \delta_{ij}} \},$$

where δ_{ij} is the censoring indicator variable defined as $\delta_{ij} = 1$ if death is observed on individual j in county i after a diagnosis of PrCA, and otherwise $\delta_{ij} = 0$.

The frailty terms W_{1i} and W_{2i} must be modeled. We refer to mutually uncorrelated county-specific effects as *spatially uncorrelated heterogeneity* and model this situation with W_{1i} and W_{2i} mutually independent Gaussian distributions according to $W_{1i} \sim i.i.d. Normal(0, \sigma_{v1}^2)$ and $W_{2i} \sim i.i.d. Normal(0, \sigma_{v2}^2)$.

As an alternative, we consider the conditional autoregressive (CAR) model, which was first introduced by Besag [26] as a smoothing algorithm in image processing. This formulation permits correlation among the random effects according to a neighborhood structure:

$$(W_{1i} | W_{1k}, \sigma_{u1}^2) \sim Normal(\bar{W}_{1i}, \sigma_{1i}^2) \\ (W_{2i} | W_{2k}, \sigma_{u2}^2) \sim Normal(\bar{W}_{2i}, \sigma_{2i}^2),$$

where

$$\begin{aligned}\bar{W}_{1i} &= \frac{1}{\sum_k \phi_{ik}} \sum_k W_{1k} \phi_{ik} \quad \text{and} \quad \bar{W}_{2i} = \frac{1}{\sum_k \phi_{ik}} \sum_k W_{2k} \phi_{ik}, \\ \sigma_{1i}^2 &= \frac{\sigma_{u1}^2}{\sum_k \phi_{ik}} \quad \text{and} \quad \sigma_{2i}^2 = \frac{\sigma_{u2}^2}{\sum_k \phi_{ik}}, \\ \phi_{ik} &= \begin{cases} 1 & \text{if regions } i \text{ and } k \text{ are adjacent, } i \neq k \\ 0 & \text{otherwise} \end{cases}.\end{aligned}$$

From the specification of the CAR distribution, it can be seen that W_{1i} and W_{2i} in the i^{th} region depend on the corresponding values in their neighborhood regions and the number of neighborhoods in the i^{th} region, hence exhibiting spatial correlation. We call W_{1i} and W_{2i} with the CAR model prior specification the *spatially correlated heterogeneity*.

We also include both spatially correlated and uncorrelated heterogeneity simultaneously in a single model to permit a trade-off between independence and a purely local spatially structured dependence of the random effects [7]. In this case, each of the terms W_{1i} and W_{2i} denotes two random effect terms. This combined spatially correlated and uncorrelated heterogeneity is called the convolution prior. In both the spatially uncorrelated heterogeneity model and the CAR model, parameters σ_{v1}^2 or σ_{u1}^2 and σ_{v2}^2 or σ_{u2}^2 control the variability of W_{1i} and W_{2i} , and need to be estimated.

3.2. Prior distribution specification and model parameter estimates

The parameter estimates for the proposed joint spatial survival model in this study are obtained through the use of a MCMC simulation method within the Bayesian framework. To conduct data analysis from the Bayesian perspective, we must specify the prior distributions for each parameter in the model. Because we have little prior information for all the parameters to be estimated, we want our data information to dominate the prior distribution by assuming reasonably non-informative priors for all parameters in this model. For all regression coefficients (β s), independent vague normal priors $N(0, 1 \times 10^6)$ were assumed (mean 0, variance 1×10^6). The shape parameters, γ_1 and γ_2 , of the survival model were given non-informative *Gamma* (1, 0.001) priors (with mean 1000, variance 1×10^6). For parameters σ_{v1}^2 or σ_{u1}^2 and σ_{v2}^2 or σ_{u2}^2 , which control the variability of the frailty effects W_{1i} and W_{2i} , the vague proper prior of *Gamma*(0.5, 0.0005) (mean 1000, variance 2×10^6) was assumed for their reciprocals (precision parameters for the random effects) [27]. When a convolution prior is specified for the random effects, there might be an identifiability problem for the two types of heterogeneity, especially in the case where a single data point is available for each small region [4, 28]. The value of the data point and the starting point of the MCMC chain for the random effects have a significant effect on the identifiability of the two types of heterogeneity [28]. To make these two types of heterogeneity identifiable in the convolution prior, the prior distribution for the precision parameters for the random effects should be chosen with caution. In some cases, an informative prior for the precision parameters may be needed. In this study, the dataset is large relative to the number of random effects to be estimated, so that an identifiability problem was not expected. Because the CAR distribution is improper and parameterized to include a sum-to-zero constraint on the random effects, a separate intercept term assigned a flat (improper uniform) prior must be included in the models [29]. Thus, when CAR models were assumed for the frailty effects W_{1i} and W_{2i} , the improper uniform prior was assumed for the intercepts α_1 and α_2 ; otherwise, the vague normal priors with 0 means were assumed for α_1 and α_2 .

WinBUGS software was used to carry out the computations in this study. As an add-on module to WinBUGS package, GeoBUGS is specially designed for fitting spatial models and producing a range of maps as output.

3.3. Model choice and adequacy assessment

The Deviance Information Criterion (DIC) [30] is adopted to guide the choice of the model used in this analysis. DIC is an intuitive extension of Akaike's Information Criterion (AIC) within the setting of Bayesian hierarchical models with random effects. The DIC is defined as

$$DIC = \bar{D} + (\bar{D} - D(\bar{\theta})) = D(\bar{\theta}) + 2(\bar{D} - D(\bar{\theta})),$$

where $\bar{D}$ is the posterior expectation of the deviance and summarizes the fit of the model, and $D(\bar{\theta})$ is the deviance evaluated at the posterior expectations of parameters. Another term p_D , defined as $p_D = (\bar{D} - D(\bar{\theta}))$, is typically used to represent the effective number of parameters. By this definition, the DIC can be considered as a Bayesian measure of fit penalized by effective number of parameters, p_D , which reflects the model complexity or degrees of freedom. If a Bayesian hierarchical model has negligible prior information, p_D will approximate the actual number of parameters and DIC will be approximately equivalent to AIC. Lower values of DIC indicate a better fitting model. It has been suggested that models with DIC values within 1 or 2 of that for the 'best' model deserve consideration, those with values within 3–7 of the 'best' model's DIC are only weakly supported, and models with a DIC more than 7 higher than that of the 'best' model are substantially inferior [30]. The computation of DIC is straightforward during an MCMC run, achieved by monitoring both θ and $D(\theta)$ in MCMC iterations. After convergence of the chain and posterior sampling, $\bar{D}$ can be estimated by the mean of the posterior samples of $D(\theta)$, and $D(\bar{\theta})$ can be estimated by plugging in the mean of the posterior samples of θ into the deviance function.

DIC measures only the relative goodness-of-fit among a collection of models. Although DIC is useful in relative identification of the best fitting model, it provides no information about the absolute adequacy of the models. Thus, other model diagnostic tools are needed to assess the adequacy of the model. We employ two diagnostic methods to assess the chosen model. One is based on the samples from the posterior predictive distribution, in which we graphically compared the Kaplan-Meier curve derived from the observed data with the curve derived from the samples from the posterior predictive distribution of the model. An overlap of the two curves might suggest an adequacy of the model [3]. The second diagnostic is based on the residuals, particularly, the Cox-Snell residuals [31]. Within the Bayesian data analysis setting, the estimated Cox-Snell residuals for each observation are calculated based on the samples from the posterior predictive distribution. If the model fitted to the observed data is satisfactory, then the estimated Cox-Snell residuals or modified Cox-Snell residuals will behave like a random sample from a unit exponential distribution. A plot of the log-cumulative hazard of the Cox-Snell residuals versus the log of Cox-Snell residuals should approximate a straight line with unit slope and zero intercept for an adequate model.

4. Application

4.1 Model choice

A set of models were fitted in our analysis. The fixed effects from covariates race, marital status and stage are always included in both the model of age at diagnosis and the model of survival time after diagnosis. For the random effects from patients' residence county,

spatially correlated and uncorrelated heterogeneity are both considered and compared. By including different types of random effects in the survival models for the two event-time outcomes, we have a pool of candidate models from which a best fitting model will be selected based on the DIC value of the joint model. In the models for (1) age at diagnosis and (2) survival time conditional on age at diagnosis, respectively, let: v_1 and v_2 denote the spatial uncorrelated heterogeneity; u_1 and u_2 denote the spatial correlated heterogeneity; u denote the common spatially correlated heterogeneity for the two outcome models; and v denote the common spatially uncorrelated heterogeneity. In summary, models considered are listed as below:

1. Model#1: v_1 , u_1 and v_2 , u_2 . This is the full model, in which the two types of heterogeneity are included simultaneously.
2. Model#2: u_1 and u_2 . Spatially correlated heterogeneity is considered in both of the two models.
3. Model#3: v_1 and v_2 . Uncorrelated heterogeneity is considered in both models.
4. Model#4: No random effects model.
5. Model#5: u_1 and v_2
6. Model#6: v_1 and u_2
7. Model#7: v
8. Model#8: u

For each of the models fitted, a single MCMC chain was run to convergence, and then a sample size of 20,000 was taken for posterior inference. Convergence was assessed graphically, by examining trace plots and time-series plots, and by checking the QQ plots of several subsets of the posterior samples. In our analysis, convergence was observed after 10,000 burn-in iterations for most of the models; for some models, convergence was slower, but after 30,000 burn-in iterations, convergence was obtained for all models. DIC values for each of the models were computed based on the 20,000 sample size after convergence of the chain was obtained.

In Table 2, we report the p_D and DIC values for the models considered. Each of the frailty models has a much smaller DIC value than the no-frailty model (Model#4). The inclusion of the frailty term in the model substantially increases the model's goodness-of-fit. The smallest DIC, and therefore the best fit, is achieved by Model#6, the model with spatially correlated and uncorrelated heterogeneity for age to diagnosis and survival after diagnosis, respectively. The full joint model with both types of heterogeneity in the two-stage survival model (Model#1) has a slightly larger DIC value compared with this best-fitting model. Further examination reveals that the posterior estimates of the main effect parameters are very close to those obtained from the best-fitting model and the full joint model. Note that in the full model, the p_D value, which measures the effective number of parameters, is much smaller than the number of random frailty parameters alone ($46 \times 4 = 184$). This implies the between-county variance is small. A further comparison of the parameter estimates from each frailty model shows the difference is very slight. To check the necessity of conditioning on age in the time to death model, an additional model based on Model#6 with the covariate age removed from the time to death model was fitted. The corresponding DIC value for this additional model, which assumes an independence of the two survival outcomes, is 116,485. This huge increase in DIC suggests that in modeling the survival time after diagnosis, the effect of age at diagnosis must be considered.

In the following section 4.2, the model adequacy assessment is conducted only for Model#6. All of the subsequent presentation of parameter estimates and interpretation in section 4.3 are also based on Model#6.

4.2 Model adequacy assessment

Figure 1 shows the Kaplan-Meier plots for checking model adequacy. In both plots, the solid line curve is based on the data itself, and the dashed line is based on the samples from the posterior predictive distribution under the best fitting model (Model#6). From Figure 1(a), it can be seen that overall the model fits the age at diagnosis data quite well, especially during the age range of 56 – 80 years marked by the two vertical lines. For age less than 56 years or age greater than 80 years, the model fit is not so good. But considering that about 83% of all the observations in our study have a diagnostic age between 56 and 80 years, the model fits most of the data well and the adequacy of the model for age at diagnosis is acceptable. The two curves in Figure 1(b) are very similar except that in the right end of the curve the estimated survival probability based on the data itself has a slightly higher value than that based on the posterior predictive samples. The overall fitting of the model to the survival data after diagnosis is good. It should be reasonable to accept this model based on the Kaplan-Meier plots.

The log Cox-Snell residual plots are shown in Figure 2. To aid visual assessment, a straight line with unit slope and zero intercept, corresponding to a perfect fitting model, is added to both plots. Although there is some deviation, the majority of points in both residual curves are close to the straight line, as is especially apparent for survival time after diagnosis in Figure 2(b). The ends of the curve marked by the two vertical lines in Figure 2(a) appear to deviate from the straight line considerably. Actually, this deviation represents only a small percentage of the total observations, i.e., these two ends of the curve have about 850 observations, only about 7% of our total observations. Further examination reveals that most of these 850 observations (about 95%) correspond to men with diagnosis age less than 55 years or age greater than 80 years. This result is consistent with that we obtained from the Kaplan-Meier plot. Hence, our model does not fit well to those data derived from subjects younger than 55 years or greater than 80 years of age. But again, considering that most observations in our data are for those with age at diagnosis between 55 and 80 years, model adequacy may be deemed reasonable. The log Cox-Snell residuals plot shown in Figure 2(b) indicates good fit for the model of survival time after diagnosis.

4.3 Parameter estimates and interpretation

The posterior means and 95% credible intervals for the model parameters in the best fitting model are given in Table 3. The posterior mean of the hazard ratio and its 95% credible interval for the main fixed-effect terms in the model also are given in Table 3. For race, and with other variables fixed, Black men have significantly higher hazards both for diagnosis of prostate cancer and for death after diagnosis of PrCA compared with Whites. Thus, for example, at each age, Black men have greater risk of PrCA diagnosis than White men, and even when diagnosed at the same age and cancer stage, Black men have increased risk of death. For marital status at diagnosis, taking single, never-married men as the referent, the other two groups of marital status have significantly lower hazards of diagnosis of prostate cancer, but the effects of marital status on survival after diagnosis are not consistent. Using the same never-married reference group, men currently married have a significantly lower hazard of death after diagnosis of prostate cancer, whereas there is no significant difference in survival after diagnosis between men who are separated, divorced or widowed and single, never-married men. The cancer stage at diagnosis is also significantly associated with the age at diagnosis and the survival after diagnosis. Compared with localized stage, regional disease is associated with a significantly younger age at diagnosis and significantly larger

hazard of death after diagnosis (shorter survival time); distant stage and unstaged cancers are associated with significantly older age at diagnosis and larger hazard of death after diagnosis. It is not surprising that the age at diagnosis has a significant effect on the survival after diagnosis: the fit reveals that older age is associated with higher hazard of death after diagnosis of prostate cancer.

The huge difference in the posterior means of the shape parameters for the two survival outcomes indicates quite different forms of hazard functions for these two survival outcomes. In modeling the age at diagnosis, the posterior mean for the shape parameter is 8.449, a value substantially larger than one, which suggests a very low and slowly increasing hazard level in early ages and a sudden, very quickly increasing hazard leading to a very high hazard in later ages. This pattern of hazard change can be seen in the estimated Kaplan-Meier survival curves in Figure 1(a). This is consistent with the fact that very few people are diagnosed with prostate cancer before 50 years of age and rates of diagnosis of prostate cancer increase until about age 75 years [19, 32], which corresponds to the age at which screening is no longer recommended [33, 34]. In the model of survival after diagnosis, the posterior mean for the shape parameter is 1.032 with a 95% credible interval of (0.993, 1.072). Since the Weibull distribution with a shape parameter of one is exactly the exponential distribution, the hazard function is constant over time. Hence, the hazard of death at any time after diagnosis is the same, irrespective of the time elapsed.

The spatially uncorrelated heterogeneity, v_1 , in the model of age at diagnosis and the spatially correlated heterogeneity u_2 in the model of survival after diagnosis were mapped as shown in Figures 3(a) and 3(b). ArcView GIS software was used to make these two maps. In order to make the spatial distribution of the random effects visually identifiable, the posterior means of the county-specific random effects were first coded according to the quintiles of their distributions. In Figure 3(a), the distribution of the spatially uncorrelated heterogeneity v_1 shows no pattern across space and hence there is no evidence for a discernible spatial pattern in the hazard for cancer diagnosis. In Figure 3(b), there is some evidence for clustering in the distribution of the spatially correlated heterogeneity u_2 . A higher hazard of death after diagnosis of PrCA appears in the central area of South Carolina. Another interesting pattern in the distribution of the spatially correlated heterogeneity u_2 is a clustering of low risk in a band-shaped area in the coastal region of the south-east part of South Carolina. This clustering pattern of low risk of death after diagnosis of prostate cancer is robust across models and independent of the type of the prior distribution chosen for the random effect from patients' residence county. In Figures 4(a) and 4(b), the maps for the distributions of county random effects on survival time after diagnosis from Model#3 and Model#5 are shown. In these two models, a spatially uncorrelated heterogeneity prior is assumed for the random effects from county on the survival after diagnosis. A spatial pattern of low risk of death after diagnosis in the coastal region of the south-east part of South Carolina is still very clearly identified in Figures 4(a) and 4(b). According to the South Carolina Cancer Facts and Figures [21, 22], the incidence rate of prostate cancer has a spatial clustering tendency in SC. There are many counties in the eastern and southern part of the state where incidence rates are significantly higher than the state average. The lower risk of death after diagnosis of prostate cancer in the south-east area of South Carolina identified in our study might be due to the local medical practice effort in prevention of prostate cancer, such as more popularity in the use of prostate-specific antigen (PSA) screening in the early detection of prostate cancer. Prostate cancer has a very long latent period and high survival rate after a diagnosis (other than in very young men) [22]. Without medical detection of prostate cancer, a man with prostate cancer might die without ever knowing he had the disease. Indeed, for most very old men this is the case. Therefore, health statistics will under-report the incidence rate of prostate cancer [22].

5. Discussion and conclusion

In this paper, we report a spatial survival analysis of prostate cancer incidence and survival in South Carolina based on population-based cancer registry and vital status data. A joint spatial survival model was implemented to explore the effects of several individual-level covariates on the two time-to-event outcomes simultaneously. The DIC criterion was used to select the best fitting model. The spatially uncorrelated and correlated heterogeneity from the frailty joint model were mapped to characterize the spatial distribution pattern for a better understanding of the geographic features of diagnosis of prostate cancer and the subsequent survival experience.

In our study, we used the DIC criterion to compare fitness across models. We found that inclusion of the frailty terms substantially improves model fit, providing evidence for correlation among outcomes from men within the same county. Inclusion of spatially uncorrelated heterogeneity for the model of age at diagnosis and spatially correlated heterogeneity for the model of survival time after diagnosis generated the best fit based on the DIC criterion. Thus, after accommodating within-county correlation among men, time to PrCA diagnosis exhibited no excess spatial correlation, whereas the model supported clusters of risk in post-diagnosis survival.

Banerjee *et al.* [4] provided another method to facilitate selection between spatially correlated and spatially uncorrelated heterogeneity. In this method, a quantity ρ is defined to represent the proportion of variability in the random effects due to clustering (spatially correlated heterogeneity): $\rho = sd(u) / (sd(u) + sd(v))$, where $sd(\cdot)$ is the empirical marginal standard deviation and u, v , respectively, denote the spatially correlated and uncorrelated heterogeneity. Evaluating ρ in our joint survival models that included both forms of heterogeneity, we found that the mean of ρ in our model for age at diagnosis was 0.201 with a 95% credible interval of (0.070, 0.481) and, in the model for survival time after diagnosis, was 0.379 with a 95% credible interval of (0.076, 0.839). These values for ρ support a spatially uncorrelated heterogeneity in the model for age at diagnosis, but do not provide definitive information for the model of survival time after diagnosis due to the wide 95% credible interval containing 0.5. This result is roughly consistent with our model choice results based on DIC, where DIC values strongly supported a spatially uncorrelated heterogeneity prior for the county random effects in modeling age at diagnosis, but the differences in DIC values across models was very small when spatially uncorrelated or correlated heterogeneity prior was used in modeling survival after diagnosis.

A possible problem with using ρ to measure the relative contribution of the two types of heterogeneity may arise from a lack of identifiability of the spatially correlated and uncorrelated heterogeneity in the convolution prior for the random effects. Our assessments for the model with convolution prior showed that, even in the case of noninformative priors, good convergence was obtained for the precision parameters for both the spatially correlated and uncorrelated heterogeneity, indicating that identifiability was not a problem in our study. Eberly and Carlin showed that despite non-identifiability of the spatially correlated and uncorrelated heterogeneity, using ρ to judge the relative presence of excess correlated and uncorrelated heterogeneity in a spatially referenced data set is sensible, because Bayesian learning about ρ can still take place [28].

One interesting result related to the spatial distribution in diagnosis and survival of prostate cancer from our study is the lack of spatial dependence in age at diagnosis but a spatially clustered pattern in survival experience following diagnosis. With the advent of PSA screening, the screening (and therefore disease detection) rates for Whites and Blacks in South Carolina are similar, even though PrCA incidence is much higher in Blacks (reflecting

real higher risk of disease) [22]. Mortality rates are very sensitive to provision of health care resources, and it is known that the poor and those living in rural areas are less likely to have good access [35, 36]. It also is known that African-American men are diagnosed with more virulent disease and at younger ages [22]. The US Southeast and SC in particular are unique in having a large population of poor, rural African Americans [37].

The identified coastal area of low hazard of death after diagnosis is robust to the type of frailty term in our model, as illustrated in Figures 4(a) and 4(b). For improved survival after diagnosis of prostate cancer, further study is necessary to identify factors that account for this cluster. Distributions of race, poverty, and access to health care resources by rurality may be a fruitful direction for future research.

As an extension to the approach advocated here, it is possible to consider both a multivariate CAR model (MVCAR) to accommodate any correlation that may exist between the spatial heterogeneity components [4], or a semi-parametric model where the hazard is based on, for example, a spline [3]. The MVCAR model may be useful when linkage between these components is of primary interest (i.e., when two different diseases are considered). Here we focus more on inference for covariates and so have not pursued that option. Another possible alternative is to consider a semi-parametric formulation for some component of the model. The hazard could be allowed to have a flexible form or the spatial component could be represented by basis functions. These options may be attractive; however, there remains the issue of how such flexible forms impact the estimation of regression parameters, as it could be found that aliasing, between, for example, spatially-defined covariates and spatially-referenced splines, could occur.

Acknowledgments

This work was supported by DOD (Department of Defense) grant GC35320342153CM through the Statewide Cancer Prevention and Control Program / Hollings Cancer Center in Charleston, SC. The authors would like to thank the South Carolina Central Cancer Registry for providing the prostate cancer data. E. Slate additionally acknowledges partial support from the NSF grant DMS-0604666, the NIH grant R01DE016353, and the NIH/NCRR COBRE project P20 RR017696.

References

1. Clayton D. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*. 1978; 65:141–151.
2. Ratcliffe SJ, Guo W, Ten Have TR. Joint modeling of longitudinal and survival data via a common frailty. *Biometrics*. 2004; 60:892–899. [PubMed: 15606409]
3. Banerjee S, Carlin BP. Semiparametric spatio-temporal frailty modeling. *Environmetrics*. 2003; 14(5):523–535.
4. Banerjee, S.; Carlin, BP.; Gelfand, AE. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC; 2004.
5. Banerjee S, Wall M, Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*. 2003; 4(1):123–142. [PubMed: 12925334]
6. Lawson, AB.; Browne, W.; Vidal-Rodeiro, CL. *Disease Mapping in WinBUGS and MLwiN*. New York: Wiley; 2003.
7. Besag J, York JC, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*. 1991; 43:1–59.
8. Tsiatis AA, Davidian MA. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*. 2004; 14:809–834.
9. Knorr-Held L, Best NG. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society Series A*. 2001; 164:73–85.

10. American Cancer Society. Cancer Facts & Figures 2005. 2005.
11. Muir CS, Nectoux J, Staszewski J. The Epidemiology of Prostatic Cancer: Geographical Distribution and Time-trends. *Acta Oncol.* 1991; 30(2):133–140. [PubMed: 2029395]
12. Farrow DC, Samet JM, Hunt WC. Regional variation in survival following the diagnosis of cancer. *J Clin Epidemiol.* 1996; 49(8):843–7. [PubMed: 8699202]
13. Helgesen F, Holmberg L, Johansson JE, Bergstrom R, Adami HO. Trends in prostate cancer survival in Sweden, 1960 through 1988: evidence of increasing diagnosis of nonlethal tumors. *J Natl Cancer Inst.* 1996; 88(17):1216–21. [PubMed: 8780631]
14. Hsing AW, Tsao L, Devesa SS. International trends and patterns of prostate cancer incidence and mortality. *Int J Cancer.* 2000; 85(1):60–7. [PubMed: 10585584]
15. Quinn M, Babb P. Patterns and trends in prostate cancer incidence, survival, prevalence and mortality. Part I: international comparisons. *BJU International.* 2002; 90(2):162–173. [PubMed: 12081758]
16. Quinn M, Babb P. Patterns and trends in prostate cancer incidence, survival, prevalence and mortality. Part II: individual countries. *BJU International.* 2002; 90(2):174–84. [PubMed: 12081759]
17. Quaglia A, Vercelli M, Puppo A, Casella C, Artioli E, Crocetti E, Falcini F, Ramazzotti V, Tagliabue G. Prostate cancer in Italy before and during the 'PSA era': survival trend and prognostic determinants. *European Journal of Cancer Prevention.* 2003; 12(2):145–152. [PubMed: 12671538]
18. Chu KC, Tarone RE, Freeman HP. Trends in prostate cancer mortality among African American men and white men in the United States. *Cancer.* 2003; 97(6):1507–16. [PubMed: 12627516]
19. Ries, LAG.; Eisner, MP.; Kosary, CL.; Hankey, BF.; Miller, BA.; Clegg, L.; Mariotto, A.; Feuer, EJ.; Edwards, BK., editors. SEER Cancer Statistics Review, 1975–2002. National Cancer Institute; Bethesda, MD: 2005.
20. American Journal of Preventive Medicine. 2006; 30(2S, Supplement):S1–S116. [PubMed: 16458783]
21. SC DHEC. South Carolina Cancer Facts and Figures, 2001–2002. 2001.
22. Drake BF, Keane TE, Mosley CM, Adams SA, Elder KT, Modayil MV, Ureda JR, Hebert JR. Prostate cancer disparities in South Carolina: Early detection, special programs, and descriptive epidemiology. *J South Carolina Med Assoc.* 2006; 102:241–9.
23. Young, JL., Jr; Roffers, SD.; Ries, LAG.; Fritz, AG.; Hurlbut, AA., editors. NIH Pub No 01-4969. National Cancer Institute; Bethesda, MD: 2001. SEER Summary Staging Manual - 2000: Codes and Coding Instructions.
24. SC DHEC, Division of Biostatistics and Public Health Informatics, Public Health Statistics and Information Services. [last updated May 18, 2006] South Carolina Community Assessment Network (SCAN) Population Information Module. <http://scangis.dhec.sc.gov/scannet/tables/populationtable.aspx>
25. SEER Stat Database: Incidence - SEER 9 Regs Public-Use, Nov 2004 Sub (1973–2002). National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch; released April 2005, based on the November 2004 submission
26. Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B.* 1974; 55:25–37.
27. Kelsall, JE.; Wakefield, JC. Discussion of “Bayesian models for spatially correlated disease and exposure data” by Best et al. In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. *Bayesian Statistics.* Vol. 6. Oxford University Press; 1999.
28. Eberly LE, Carlin BP. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine.* 2000; 19:2279–2294. [PubMed: 10960853]
29. Besag J, Kooperberg CL. On conditional and intrinsic autoregressions. *Biometrika.* 1995; 82:733–746.
30. Spiegelhalter DJ, Best N, Carlin BP, Van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B.* 2002; 64:583–639.
31. Collett, D. Modeling Survival Data in Medical Research. London: Chapman and Hall; 1994.

32. US Cancer Statistics Working Group. United States Cancer Statistics: 2002 Incidence and Mortality. Atlanta (GA): US DHHS/CDC/NIH-NCI; 2005.
33. Gosselaar C, Roobol MJ, Schroder FH. Prevalence and characteristics of screen-detected prostate carcinomas at low prostate-specific antigen levels: aggressive or insignificant? *BJU International*. 2005; 95(2):231–7. [PubMed: 15667646]
34. Schmid H-P, Riesen W, Prikler L. Update on screening for prostate cancer with prostate-specific antigen. *Crit Rev Oncol Hematol*. 2004; 50(1):71–8. [PubMed: 15094160]
35. Probst JC, Moore CG, Glover SH, Samuels ME. Person and place: the compounding effects of race/ethnicity and rurality on health. *Am J Public Health*. 2004; 94(10):1695–703. [PubMed: 15451735]
36. Larson SL, Fleishman JA. Rural-urban differences in usual source of care and ambulatory service use: analyses of national data using Urban Influence Codes. *Med Care*. 2003; 41(7 Suppl):III65–III74. [PubMed: 12865728]
37. Hebert JR, Elder K, Ureda JR. Meeting the challenges of cancer prevention and control in South Carolina: Focus on seven cancer sites, engaging partners. *J South Carolina Med Assoc*. 2006; 102:177–82.

Figure 1.

Empirical Kaplan-Meier curves to assess goodness-of-fit for the models of: (a) age at diagnosis; and (b) survival time after a diagnosis. The solid line is the curve based on the data itself, and the dashed line (with 95% pointwise credible limits) is the curve based on the samples from the posterior predictive distribution under the best fitting model.

Figure 2.

Log Cox-Snell residual plots to assess goodness-of-fit for the models of: (a) age at diagnosis; and (b) survival time after a diagnosis. The straight line in each plot has unit slope and zero intercept.

Figure 3.

Map of South Carolina showing: (a) spatially uncorrelated heterogeneity, v_1 , in the best fitting model for age at diagnosis; and (b) spatially correlated heterogeneity, u_2 , in the best fitting model for survival after diagnosis.

Figure 4.

Map of South Carolina showing spatially uncorrelated random effects in: (a) joint Model#3; and (b) joint Model#5 for survival after diagnosis, demonstrating similar spatial patterns of survival relative to those obtained in the best fitting model (see Figure 3b).

Table 1

Characteristics of prostate cancer patients; South Carolina, 1997–2001.

Covariate	N	Percent (%)
Race		
Black/African American	3,874	31.66
White/European American	8,364	68.34
Marital status at diagnosis		
Single(never married)	823	6.73
Married	9,518	77.77
Separated, divorced, widowed	1,897	15.50
Cancer stage at diagnosis		
Localized	9,480	77.46
Regional	1,391	11.37
Distant	618	5.05
Unstaged	749	6.12
Vital status as of 12/31/2003		
Alive	9,989	81.62
Dead	2,249	18.38

Table 2

p_D and DIC values for the candidate models.

Model	p_D	DIC
Model#1	55.9	115724
Model#2	55.0	115735
Model#3	55.7	115725
Model#4	16.8	115845
Model#5	57.3	115733
Model#6	57.0	115722
Model#7	41.2	115768
Model#8	41.7	115762

Table 3

Parameter estimate from the posterior distribution in the best-fitted model.

Covariates	Posterior Mean (95% Credible interval)	Hazard Ratio (95% Credible interval)
<u>Model of the age at diagnosis</u>		
Intercept	-35.810 (-36.290, -35.340)	
Race (White/ European American*)		
African American	0.094 (0.054, 0.135)	1.098 (1.054, 1.143)
Marital status at diagnosis(single)		
Married	-0.166 (-0.239, -0.093)	0.848 (0.790, 0.909)
separated, divorced, widowed	-0.591 (-0.673, -0.508)	0.554 (0.511, 0.600)
Cancer Stage at diagnosis(localized)		
Regional	0.263 (0.207, 0.321)	1.302 (1.230, 1.378)
Distant	-0.570 (-0.653, -0.487)	0.566 (0.520, 0.614)
Unstaged	-0.722 (-0.801, -0.646)	0.486 (0.450, 0.524)
Shape parameter γ_1	8.449 (8.342, 8.561)	
σ_{v1} -- spatially uncorrelated heterogeneity	0.105 (0.078, 0.140)	
<u>Model of survival after diagnosis</u>		
Intercept	-10.650 (-11.060, -10.250)	
Race (white)		
African American	0.218 (0.126, 0.311)	1.254 (1.141, 1.378)
Marital status at diagnosis(single)		
Married	-0.191 (-0.349, -0.030)	0.825 (0.703, 0.972)
separated, divorced, widowed	0.074 (-0.096, 0.244)	1.080 (0.905, 1.280)
Cancer stage at diagnosis(localized)		
Regional	0.238 (0.089, 0.381)	1.263 (1.090, 1.454)
Distant	1.862 (1.749, 1.975)	6.459 (5.756, 7.241)
Unstaged	0.670 (0.532, 0.802)	1.956 (1.705, 2.224)
Age at diagnosis	0.068 (0.064, 0.074)	1.070 (1.065, 1.075)
Shape parameter γ_2	1.032 (0.993, 1.072)	
σ_{u2} -- spatially correlated heterogeneity	0.194 (0.088, 0.330)	

---

*The groups enclosed in the parenthesis are the reference groups for each categorical explanatory variable