

# Bayesian modeling of clustered competing risks survival times with spatial random effects

Somayeh Momenyan <sup>(1)</sup>, Amir Kavousi <sup>(2)</sup>, Taban Baghfalaki <sup>(3)</sup>, Jalal Poorolajal <sup>(4)</sup>

(1) Department of Biostatistics, Shahid Beheshti University of Medical Sciences, Tehran, Iran

(2) Workplace Health Promotion Research Center and Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran

(3) Department of Statistics, School of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran

(4) Research Center for Health Sciences and Department of Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran

**CORRESPONDING AUTHOR:** Amir Kavousi, Department of Biostatistics, Shahid Beheshti University of Medical Sciences, Tehran, Iran. E-mail: kavousi@sbmu.ac.ir

**DOI:** 10.2427/13301

Accepted on May 14, 2020

## ABSTRACT

In some studies, survival data are arranged spatially such as geographical regions. Incorporating spatial association in these data not only can increase the accuracy and efficiency of the parameter estimation, but it also investigates the spatial patterns of survivorship. In this paper, we considered a Bayesian hierarchical survival model in the setting of competing risks for the spatially clustered HIV/AIDS data. In this model, a Weibull Parametric distribution with the spatial random effects in the form of county-failure type-level was used. A multivariate intrinsic conditional autoregressive (MCAR) distribution was employed to model the areal spatial random effects. Comparison among competing models was performed by the deviance information criterion and log pseudo-marginal likelihood. We illustrated the gains of our model through the simulation studies and application to the HIV/AIDS data.

*Key words:* Proportional hazards model; Competing risks; Spatial random effect; Markov chain Monte Carlo; HIV; AIDS.

## INTRODUCTION

In biomedical studies, it is common to have time to event data. In the survival analysis, in many situations, there are some risk factors that are unobservable. In the presence of such risk factors, the usual survival models such as the Cox proportional hazards model are not proper (1). For solving this problem, Vaupel et al. introduced a model with the random effect (2). Clayton and Cuzik introduced the generalization of proportional hazards model by including a random effect to the Cox proportional hazards model to account for variability (3). Frailty models could increase the accuracy and efficiency of the parameter estimation when survival data are independent.

But, in some situations, survival data are dependent such as when a sample of individuals was grouped into clusters. If individuals come from different areas, there will be a spatial correlation between survival data because data from the same or nearer areas are expected to be more similar than those from farther areas (4).

In biostatistics and epidemiology studies, modeling spatial survival data accounting for the spatial association has become popular. For example, Li and Ryan (5) and Banerjee et al. (6) proposed a spatial survival model using the proportional hazards structure from the classical and Bayesian perspective, respectively. The survival model for taking spatiotemporal variation in survival data was investigated by Banerjee and Carlin (6) and Hanson et al. (7). Banerjee and Dey presented semiparametric hierarchical modeling for the proportional odds model in spatially survival data (8). Diva et al. used a Bayesian hierarchical survival model within the proportional hazards and the proportional odds settings (9). Pan et al. proposed a spatial Bayesian semiparametric model to analyze interval-censored survival data (10). More recently, Cramb et al. proposed a spatial flexible

parametric relative survival model (11), and Zhou and Hanson applied the semiparametric survival model with spatial random effects to arbitrarily censored survival data (12).

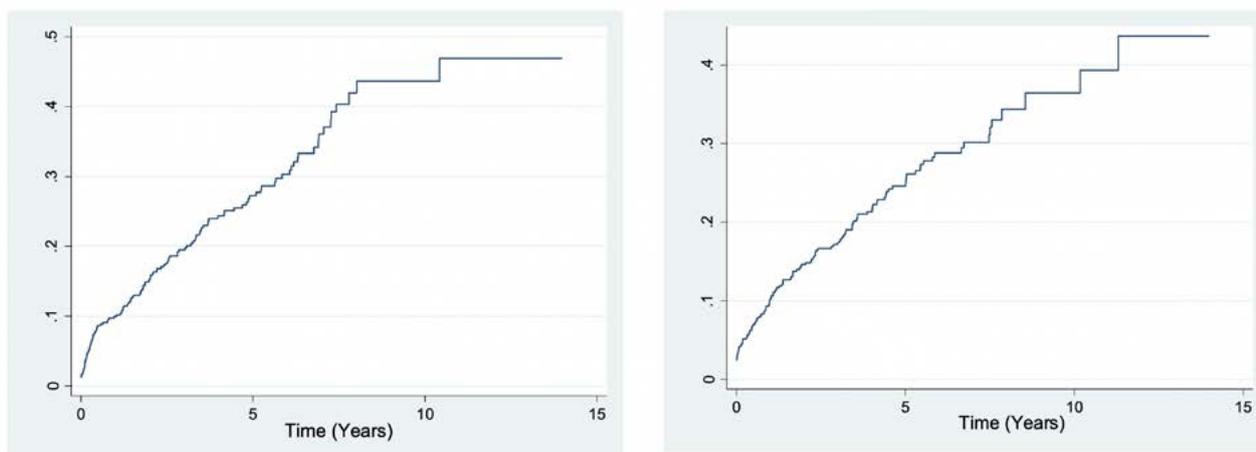
In the survival data, there is also a situation where, there is more than one cause of failure, but only the occurrence of the first one is observable. (13). The new aspect of this paper is an extension of survival model from single failure type to competing risks in spatially sampled data.

The rest of this paper is outlined as follows. Section 2 describes the HIV/AIDS data. Section 3 formulates our proposed model, including the notations, modeling the cause-specific proportional hazards function with spatial random effects, the prior distribution for areal spatial random effects, and the prior and posterior specification. Section 4 discusses a Bayesian model comparison using deviance information criterion (DIC) (5) and log pseudo-marginal likelihood (LPML) (14). In Section 5, the performance of our model is evaluated through simulation studies. Section 6 analyzes the HIV/AIDS data. Finally, Section 7 presents a discussion of related issues.

## THE HIV/AIDS DATA

The data were from a retrospective cohort study, which was conducted in Hamadan Province, the central-western part of Iran, from 1997 to 2011. All 585 HIV-positive people who had a medical record in the HIV testing and treatment centers were included in this study. The explanatory variables included were as follows: age at time of diagnosis, sex, marital status, method of transmission, co-infection with tuberculosis (TB). Also, date of HIV diagnosis, date of progression to AIDS and date of death (if any) and patient's county of residence were collected. The main outcome in this study was the time

**FIGURE 1. Cumulative incidence function for AIDS progression (left) and mortality post-HIV infection (right)**



interval, between HIV diagnosis and AIDS progression, so the event of interest was AIDS progression. However, some of HIV infected patients die before AIDS progression. Thus, death before AIDS was considered as competing risk. The patients who did not experience any of these events or were lost to follow-up were considered as censored. Accordingly, in the final outcome classification, patients were categorized into three categories: those who developed AIDS (23.4%), those who died post-HIV infection (22.9%), and those who censored (53.7%). Figure 1 plotted the observed cumulative incidence curve for risks of AIDS progression and mortality post-HIV infection. As seen, the cumulative incidence probability for AIDS progression is higher than competing event of death after the first 5 years.

### MODEL FORMULATION

In this section, we suppose  $n$  individuals come from  $K$  counties.  $n_k$  is the number of individuals in a sample in each county, where  $k = 1, \dots, K$  and  $\sum_{k=1}^K n_k = n$ . Also, let  $C_{ik} = (T_{ik}, \delta_{ik})$  be the data for the  $i$ -th individual living in the  $k$ -th county,  $i = 1, \dots, n_k$ , where  $T_{ik}$  denotes time to an event which may be right censored time. In competing risks data, each individual has one of the  $G$  possible failure types during follow-up or could be right censored. Hence, failure type indicator or  $\delta_{ik}$  takes value from  $\{0, 1, \dots, G\}$ .

#### Survival model with random effects

Different models for analyzing competing risks data have been proposed in the last three decades [15, 16]. The typical approach for analyzing competing risks survival times is modeling the cause-specific hazard functions [17].

$$h_{ik}^g(t|x_{ik}) = \lim_{dt \rightarrow 0} \frac{p(t \leq T_{ik} < t+dt, \delta_{ik} = g | T_{ik} \geq t, x_{ik})}{dt} = h_0(t)^g \exp(x_{ik}^T \beta^g), \tag{1}$$

where  $h_0(t)^g$  is an unspecified baseline risk function and  $\beta^g = (\beta_1^g, \beta_2^g, \dots, \beta_p^g)$  is a  $p \times 1$  vector of regression parameters associated to a  $p \times 1$  vector of observed explanatory covariates  $x_{ik} = (x_{ik1}, x_{ik2}, \dots, x_{ikp})$  for the  $g$ th type of failure where  $g = 1, \dots, G$ . In this model, to estimate the hazard of each failure type, other competing risks are assumed as censored in addition to those who are censored from loss to follow-up. But, sometimes, those who experience competing risk events are censored informatively; because of that, in the competing risks setting, the assumption of independent censoring is no longer reasonable. For example, in the HIV/AIDS data, the factors that affect the risk of AIDS progression as the event of interest might also influence the probability of mortality post-HIV infection as the competing risk event. To avoid biased results, we introduce the random effects  $V_{ik} = (V_{ik}^1, \dots, V_{ik}^G)^T$ , to capture

for the association between times of different failure types; for more details, see Huang and Wolfe [18] and also Christian et al. [19]. We introduce these random effects through (1), as

$$h_{ik}^g(t|x_{ik}, v_{ik}^g) = h_0(t)^g \exp(x_{ik}^T \beta^g + v_{ik}^g) \tag{2}$$

On the other hand, since subjects come from different counties, we consider a survival model with the spatial random effects. Let  $W_k^g, k = 1, \dots, K, g = 1, \dots, G$ , denotes the spatial random effects. We introduce again, these random effects through (2), as

$$h_{ik}^g(t|x_{ik}, v_{ik}^g, W_k^g) = h_0(t)^g \exp(x_{ik}^T \beta^g + v_{ik}^g + W_k^g) \tag{3}$$

Alternatively, for the baseline hazard function, we assume a cause-specific Weibull function, which is  $h_0(t)^g = p^g t^{p^g-1}$  with the shape parameter  $p^g$  for the  $g$ th type of failure. Hence, the model (3) can be substituted by:

$$h_{ik}^g(t|x_{ik}, v_{ik}^g, W_k^g) = p^g t^{p^g-1} \exp(x_{ik}^T \beta^g + v_{ik}^g + W_k^g) \tag{4}$$

#### Spatial random effects

A common model for areal data collected over a geographic region in a univariate case such as a single disease is the conditional autoregressive (CAR) distribution, developed by Besage [20]. Let  $W = (W_1, \dots, W_K)^T$  be the spatial random effects vector observed at the  $K$  areal counties, then the general form of the CAR joint distribution is

$$W \sim N(0, [\sigma_w^2 D (I - \alpha B)]^{-1}) \tag{5}$$

where  $B$  is a  $K \times K$  matrix,  $\sigma_w^2$  is conditional variance and  $\alpha$  is called a smoothing parameter. If we specify  $D = \text{diag}(m_k), m_k$  is the number of neighbors of the  $k$ -th county, and  $B = D^{-1} B_w$  where  $B_w$  is the adjacency matrix of the graph representing our county ( $b_{wkk} = 0, b_{wkk} = 1$  if the county  $k$  is a neighbor of the county  $k, b_{wkk} = 0$  if the county  $k$  is not a neighbor of the county  $k$ ), we obtain the so-called intrinsic conditional autoregressive (ICAR) distribution that is the most common CAR distribution [21]. Also, in this structure, we set the smoothing parameter  $\alpha = 1$ . This distribution is denoted as  $CAR(1, \sigma_w^2)$  and its formulation, thus becomes

$$W \sim N(0, \sigma_w^2 \Sigma_w^{-1}), \tag{6}$$

where  $\Sigma_w = (D - B_w)$  is a  $K \times K$  matrix. But, when we have the multivariate areal data, such as information on multiple diseases over the same regions, the multivariate areal models are proper for analyzing this kind of data [22, 23]. There are several multivariate areal models that have been proposed in the literature. The multivariate Normal Markov random field (MRF) was proposed by

Mardia (24). A twofold CAR model for two diseases and multi-objective version of the CAR model were proposed by Kim et al. (25) and sain and Cressie (26), respectively. The multivariate CAR models for the hierarchical modeling based on MRF have been proposed by Carlin and Banerjee (27) and Gelfand and Vounatsou (28). In this paper, we use the multivariate intrinsic conditionally autoregressive (MCAR) distribution for the spatial random effects in the HIV/AIDS data to be able to model the possible correlation in the risks of AIDS progression and mortality post-HIV infection over the same counties as well as the possible spatial correlation among the counties. Let  $\mathbf{W} = (W_1^T, \dots, W_K^T)^T$  where  $\mathbf{W}$  is  $KG \times 1$  with each  $W_k = (W_k^1, \dots, W_k^G)^T$  being a  $G$ -dimensional vector of the spatial random effects collection for the  $G$  possible failure types within the  $k$ -th county. The joint distribution for  $\mathbf{W}$  takes the following form

$$\mathbf{W} \sim N(\mathbf{0} [\mathbf{D} (\mathbf{I} - \alpha \mathbf{B}), \otimes \Lambda]^{-1}). \quad (7)$$

If we postulate  $\mathbf{D}$ ,  $\mathbf{B}$ , and  $\alpha$  similar to ICAR distribution that was mentioned above, we obtain the multivariate intrinsic conditional autoregressive distribution. This distribution is denoted as MCAR ( $\mathbf{I}, \Lambda$ ) and its formulation is also equivalent to

$$\mathbf{W} \sim N(\mathbf{0}, (\Sigma_w \otimes \Lambda)^{-1}), \quad (8)$$

where  $\Sigma_w = (\mathbf{D} - \mathbf{B}_w)$  is a  $K \times K$  matrix and therefor  $\Sigma_w \otimes \Lambda$  is a  $KG \times KG$  matrix. Moreover,  $\Lambda$  is a  $G \times G$  positive definite and symmetric matrix, which is defined as the variance-covariance matrix of the spatial failure type and controls the correlation between the competing risks in the survival process at any given county. Note that, in this setup, we are modeling the dispersion matrix of the spatial-by-failure type interaction,  $\mathbf{W}$ , as the Kronecker product of the spatial dispersion and the failure type dispersion. Also, we use the same  $\Lambda$  for all of the counties.

### Likelihood

Let  $\theta = (\beta^g, \Lambda, W_k^g, v_{ik}^g, p^g; i = 1, \dots, n_k, k=1, \dots, K, g=1, \dots, G)$  denotes all the unknown parameters and  $(C_{ik}, X_{ik}; i=1, \dots, n_k, k=1, \dots, K)$  represents the observed data; therefore, the likelihood function for this model is product of all  $n$  individual contributions to the likelihood

$$L(\theta | C, X) \propto \prod_{k=1}^K \prod_{i=1}^{n_k} L_{ik}(\theta | C_{ik}, X_{ik}) \quad (9)$$

Considering the distribution of competing risks survival response,  $L_{ik}(\theta | C_{ik}, X_{ik})$  is as follows:

$$\prod_{g=1}^G \left[ [p^g t^{p^g-1} \exp(x_{ik}^T \beta^g + v_{ik}^g + W_k^g)]^{I(\delta_{ik}=g)} \times \exp \left[ - \int_0^{t_{ik}} p^g t^{p^g-1} \exp(x_{ik}^T \beta^g + v_{ik}^g + W_k^g) dt \right] \right] \quad (10)$$

Finally, the likelihood for  $\theta$ , conditional on the observed data is expressed as follows,

$$\prod_{k=1}^K \prod_{i=1}^{n_k} \left[ \prod_{g=1}^G [p^g t^{p^g-1} \exp(x_{ik}^T \beta^g + v_{ik}^g + W_k^g)]^{I(\delta_{ik}=g)} \times \exp \left[ - \int_0^{t_{ik}} p^g t^{p^g-1} \exp(x_{ik}^T \beta^g + v_{ik}^g + W_k^g) dt \right] \right] \quad (11)$$

### Bayesian approach

In the Bayesian approach, we considered the prior distribution for all parameters. The standard non-informative prior distributions for the parameters were considered as follows: for the regression coefficients,  $\beta^g$ , and the Weibull shape parameters,  $p^g; g = 1, \dots, G$ ; a multivariate Normal,  $N(\mathbf{0}, \Sigma_\beta)$ , and a gamma,  $\mathcal{G}(a_1, b_1)$ , priors were taken respectively. We assumed for  $v_{ik}^g; i = 1, \dots, n_k, k = 1, \dots, K, g = 1, \dots, G$  a multivariate Normal distribution, that is  $V_{ik} \sim N(\mathbf{0}, \Lambda)$  where  $\Lambda$  is a  $G \times G$  variance-covariance matrix. Concerning the spatial random effects, we used the MCAR prior distribution,  $\mathbf{W} \sim \text{MCAR}(\mathbf{I}, \Lambda)$ , (see section 3.2). For the variance-covariance matrix,  $\Lambda$  an inverse Wishart,  $JW(\mathbf{R}, \nu)$ , was used where matrix  $\mathbf{R}$  is a  $G \times G$  positive definite matrix and  $\nu$  is degrees of freedom. Thus, in our model, we are using the same  $\Lambda$  to model the MCAR and failure type random effects distributions. For all these priors, we chose the distributions with the very high variance. The joint posterior distribution of our proposed model is denoted by  $\pi(\theta | C, X)$ , and is proportional to

$$L(\theta | C, X) \times \pi(\beta^g | \Sigma_\beta) \times \pi(p^g | a_1, b_1) \times \pi(V, \Lambda) \times \pi(W | \Sigma_w, \Lambda) \times \pi(\Lambda | R, \nu). \quad (12)$$

After initializing values for the parameters, sampling from the full conditional distribution by the MCMC algorithm was performed. All statistical analysis and also mapping the results were performed using OpenBUGS software (29), version 3.2.3, GeoBUGS and R package R2OpenBUGS (30, 31).

### MODEL SELECTION

We selected two summary measures for the model selection as follows: the deviance information criterion (DIC) and log pseudo-marginal likelihood (LPML). The DIC criterion for model  $m$  is defined as follows:

$$DIC(m) = \overline{-2D[\theta_m, m]} - D[\overline{\theta_m}, m] = D[\overline{\theta_m}, m] + 2p_m, \tag{13}$$

where  $D[\theta_m, m]$  is the deviance measure and is defined as:

$$D[\theta_m, m] = -2\log f(Y|\theta_m, m),$$

and  $\overline{D[\theta_m, m]}$  is its posterior mean. Also,  $\overline{\theta_m}$  is the posterior mean of the parameters involved in model  $m$  and  $p_m$  is an effective number of parameters of model  $m$  and is defined as follows:

$$p_m = \overline{D[\theta_m, m]} - D[\overline{\theta_m}, m].$$

The model with the smallest of DIC is the preferred model between a collection of alternative models. LPML statistic is based on the conditional predictive ordinate (CPO). The CPO statistic for  $ik$ th individual is defined as predictive density based on all of data except  $ik$ th individual. The high value of CPO statistic indicates that data for subject can be truly predicted by a model based on data from all other subjects. The LPML is used as an overall measure of the model fit and is calculated as:

$$LPML = \sum \log(CPO_{ik}). \tag{14}$$

A model with a maximum value of LPML implies a better model. Using the MCMC methods, the DIC and LPML compute from the posterior samples.

## SIMULATION STUDY

The performance of our model was evaluated through a series of scenarios in the simulation study. We generated a total of 100 simulated datasets with three levels of sample size in each county ( $n_k=50, n_k=100, n_k=200$ ) and with three censoring rate (low, 20%, medium, 40%, and high, 60%). For each dataset, the neighborhood structure was based on the nine counties in Hamadan, Iran. We used a continuous covariate ( $X_1$ ) and a binary categorical covariate ( $X_2$ ) that was generated from a standard Normal distribution and a Bernoulli distribution with a success probability  $p = 0.5$ , respectively. Also, we simulated two competing risks, risk 1 and 2, from the cause-specific exponential model. The failure times were generated based on the algorithm that Beyersmann and et al. presented for Competing risks data (32). As for each individual, the event time was generated by overall hazard rate that at each time point was the sum of cause-specific hazard rates for two types of event,  $\lambda_{ik}(t) = \lambda_{ik}^1(t) + \lambda_{ik}^2(t)$ . The cause-specific hazard rate for each type of event was considered as follows in each county

$$\lambda_{ik}^1(t|x_{ik}, V_{ik}^1) = \exp(\beta_1^1 x_{1ik} + \beta_2^1 x_{2ik} + V_{ik}^1 + W_k^1),$$

$$\lambda_{ik}^2(t|x_{ik}, V_{ik}^2) = \exp(\beta_1^2 x_{1ik} + \beta_2^2 x_{2ik} + V_{ik}^2 + W_k^2).$$

The type of event was determined by a Bernoulli experiment with the probability  $P_1 = \lambda^1 / \lambda$  for the event of type 1 and  $P_2 = \lambda^2 / \lambda$  for the event of type 2. The random effects  $V_{ik}$  were generated from a multivariate Normal with a mean 0 and a  $2 \times 2$  covariance matrix of  $\Lambda$ , that  $\Lambda_{11} = 0.1, \Lambda_{22} = 0.1$  and  $\Lambda_{12} = 0.05$  and then were centered around its mean. Also, the spatial random effects  $W$ , were generated from the MCAR distribution,  $N(0, (\Sigma_w^{-1} \otimes \Lambda))$  where  $\Sigma_w^* = 0.99 \times \Sigma_w + \text{diag}(0.01)$ , and then were centered around its mean. We introduced  $\Sigma_w^*$  to be invertible the precision matrix and centered to be identifiable the spatial random effects. The corresponding regression coefficients were set  $\beta_1^1 = 0.6, \beta_2^1 = -0.4$  for risk 1 and  $\beta_1^2 = -0.3, \beta_2^2 = 0.7$  for risk 2. The matrix  $R$  was set as  $0.1I_{2 \times 2}$ . The data above a threshold were right censored which were selected as the  $(1-\alpha)$ -quantile of the sample survival times, such that 100 $\alpha$ % of the observations were censored.

For each of the fitted models, the convergence of the MCMC chain was evaluated by trace plot, auto-correlation plot, and Gelman-Rubin's diagnostic (33). For each MCMC chain, we run the 25,000 iterations. For each parameter, the point estimate, bias, and mean square error (MSE) were calculated by the average of the means, biases and the square errors from the 100 replicates, respectively. The coverage probability (CP) was calculated as the proportion of the 95% credible intervals that contain the true values. The bias and square error are defined as:

$$\text{Bias}(\theta) = (\hat{\theta}_i - \theta), \quad \text{SE}(\theta) = (\hat{\theta}_i - \theta)^2.$$

## Simulation results

The results of the simulation study from nine scenarios are reported in Table 1. While estimation of the regression coefficients in the total scenario were close to the truth value 1, but estimation of the covariance matrix  $\Lambda$  had a consistently small bias for the large sample size ( $n_k = 200$ ) and with censoring rate 20% and 40%. Also, all parameters had the coverage probabilities close to the nominal level of 0.95 under all the scenarios. The MSE criterion for all parameters was close together and as sample size increases, the estimation accuracy of parameters increases.

## ANALYSIS OF THE HIV/AIDS DATA

We analyzed the HIV/AIDS data using four models as follows:

Model 1: The model with any random effects in the log-relative hazard ( $x_{ik}^T \beta^g$ ),

**TABLE 1. Estimation results for the simulation study**

Censoring rate		$n_k = 50, n = 450$				$n_k = 100, n = 900$				$n_k = 100, n = 1800$			
		Estimate	Bias	MSE	CP	Estimate	Bias	MSE	CP	Estimate	Bias	MSE	CP
<b>20%</b>													
	$\beta_1^1 = 0.6$	0.598	-0.002	0.019	96	0.594	-0.006	0.009	95	0.584	-0.016	0.005	97
	$\beta_1^2 = -0.3$	-0.276	-0.024	0.017	94	-0.287	-0.013	0.009	95	-0.294	-0.006	0.004	94
	$\beta_2^1 = -0.4$	-0.390	-0.010	0.028	96	-0.385	-0.015	0.013	98	-0.389	-0.011	0.007	92
	$\beta_2^2 = 0.7$	0.682	-0.018	0.028	95	0.698	-0.002	0.015	93	0.690	-0.010	0.006	95
	$\Lambda_{11} = 0.1$	0.085	-0.015	0.004	99	0.093	-0.007	0.003	97	0.097	-0.003	0.002	98
	$\Lambda_{22} = 0.1$	0.084	-0.016	0.004	99	0.091	-0.009	0.002	99	0.093	-0.007	0.001	99
	$\Lambda_{12} = 0.05$	0.031	-0.019	0.001	100	0.043	-0.007	0.0002	99	0.044	-0.006	0.0001	99
<b>40%</b>													
	$\beta_1^1 = 0.6$	0.595	-0.005	0.022	97	0.613	0.013	0.010	96	0.610	0.010	0.005	94
	$\beta_1^2 = -0.3$	-0.285	-0.015	0.023	96	-0.306	0.006	0.011	94	-0.290	-0.010	0.005	95
	$\beta_2^1 = -0.4$	-0.397	-0.003	0.039	95	-0.396	-0.004	0.019	95	-0.393	-0.007	0.009	95
	$\beta_2^2 = 0.7$	0.704	0.004	0.039	95	0.708	0.008	0.017	97	0.691	-0.009	0.008	97
	$\Lambda_{11} = 0.1$	0.087	-0.013	0.006	99	0.090	-0.010	0.004	98	0.090	-0.010	0.003	95
	$\Lambda_{22} = 0.1$	0.080	-0.020	0.007	97	0.086	-0.014	0.006	100	0.087	-0.013	0.002	98
	$\Lambda_{12} = 0.05$	0.021	-0.029	0.003	96	0.030	-0.020	0.0005	99	0.053	0.003	0.0003	99
<b>60%</b>													
	$\beta_1^1 = 0.6$	0.642	0.042	0.030	95	-0.586	-0.014	0.013	95	0.594	-0.006	0.005	98
	$\beta_1^2 = -0.3$	-0.302	0.002	0.025	97	-0.303	0.003	0.015	92	-0.284	-0.016	0.006	93
	$\beta_2^1 = -0.4$	-0.368	-0.032	0.055	97	-0.339	-0.061	0.026	96	-0.395	-0.005	0.012	96
	$\beta_2^2 = 0.7$	0.749	0.049	0.071	93	0.670	-0.030	0.030	96	0.696	-0.004	0.015	90
	$\Lambda_{11} = 0.1$	0.082	-0.018	0.008	100	0.089	-0.011	0.006	98	0.090	-0.010	0.005	100
	$\Lambda_{22} = 0.1$	0.078	-0.022	0.010	99	0.083	-0.017	0.009	99	0.087	-0.013	0.004	99
	$\Lambda_{12} = 0.05$	0.020	-0.030	0.009	95	0.022	-0.028	0.002	98	0.026	-0.024	0.001	93

**TABLE 2. Model comparison results for the HIV/AIDS data, bold number shows the best fit**

	Log-relative hazard	LPML	DIC
<b>(Model 1)</b>	$x_{ik}^T \beta^g$	-963.11	1908.00
<b>(Model 2)</b>	$x_{ik}^T \beta^g + W_k$	-962.60	1906.00
<b>(Model 3)</b>	$x_{ik}^T \beta^g + V_{ik}^g + W_k^g$	-962.76	1906.00
<b>(Model 4)</b>	$x_{ik}^T \beta^g + V_{ik}^g + W_k^g$	-958.78	1903.00

$$(x_{ik}^T \beta^g + W_k; W \sim N(0, \Sigma_w^{-1})),$$

Model 3: The model with the spatial and non-spatial failure type random effects in the log-relative hazard

$$(x_{ik}^T \beta^g + V_{ik}^g + W_k^g; W \sim N(0, \mathbf{I}_{K \times K}^{-1} \otimes \Lambda)),$$

Model 4: The model with the multivariate spatial and failure type random effects in the log-relative hazard

$$(x_{ik}^T \beta^g + V_{ik}^g + W_k^g; W \sim N(0, \mathbf{I} \Sigma_w^{-1} \otimes \Lambda)).$$

The first model is the model that does not account for the correlation between the competing risks in the survival process and also the spatial correlation among the counties. The second model is the model that only incorporates the spatial variation among the counties. In this model, we considered the ICAR prior distribution for the univariate spatial random effects (see section 3.2) and also for the spatial random effects variance  $\sigma_w^2$  an inverse-gamma,  $IG(a_2, b_2)$  prior was used. The third model is the spatially independent model with basically the same structure with our proposed model. The spatially independent model was accomplished by replacing the  $\Sigma_w$  with a diagonal matrix which does not account for the spatial structure of the counties, whence the collection of independent random effects for  $G$  failure types within the county, denoted by  $W$  has a  $N(0, (\mathbf{I}_{K \times K}^{-1} \otimes \Lambda))$  distribution. In other words, in this model, we consider the spatial and non-spatial failure type random effects in the log-relative hazard. The fourth model is the full model that we proposed in this paper.

The proposed models were fitted based on sampling chains of 100,000 iterations with a spacing of 10 for reducing level of correlation and the first 10,000 discarded as a burn-in. Trace, auto-correlation and density plots of the posterior distributions were assessed for convergence of the MCMC chains. The Gelman-Rubin's statistics for all parameters were between 1.0 and 1.1. The consistent batch means estimates of Monte Carlo standard errors were from 0.00002 to 0.0002 for all parameters. The summary measures of all parameters consist of the mean and standard deviation by posterior samples were obtained. Moreover, for the regression coefficients, the adjusted hazard ratios (HR) were calculated with 95% credible intervals.

Table 2 compared four models based on two criteria, DIC and LPML. The DIC and LPML values for the four proposed models were very close, but model 4 which introducing the multivariate spatial and failure type random effects, showed somewhat better comparison measure values. In addition, these values suggested that each of these models is better than the first model that does not account for any correlation. The summary measures of all parameters of the fourth model were presented in Table 3. Also, because the MCAR distribution is improper and parameterized to include a sum-to-zero constraint on the

random effects, a separate intercept coefficient for each of two risks with a flat prior was included in this model.

Based on the adjusted HR estimates, there were the significant relationships between TB co-infection and mode of transmission with the risk of AIDS progression. However, the adjusted relationships for the risk of mortality post-HIV infection were not statistically significant for all the predictors except age at diagnosis. In other words, HIV-positive patients who were co-infected with TB and became infected through mother to child had a higher risk of AIDS progression as compared to those who were infected with HIV alone and infected through IDU. Also, the risk of mortality post-HIV infection was higher in patients aged 45 to 74 years than in those aged 0 to 24 years, such that the adjusted HR was 2.33.

We also mapped the summaries of our results. Figure 2 shows 2 maps that represent the posterior spatial relative risk in nine counties of Hamadan Province. The map on the left is for the relative risk of AIDS progression and the map on the right is for the relative risk of mortality post-HIV infection that is defined by  $\exp(W_w^1)$  and  $\exp(W_w^2)$  respectively. The posterior estimates of county-specific random effects were recorded based on the quintile of their distribution for showing the spatial inequalities on the map. As shown in Figure 2, for the risk of AIDS progression, one cluster of counties was identified with higher risk in the south, southeast, and southwest regions (six out of nine counties) and one cluster with lower risk in the north, northeast, and northwest regions (three out of nine counties) was identified. Also, for the risk of mortality post-HIV infection, the high-risk cluster consists of the northwest, west, and southwest regions (five out of nine counties) and the low-risk cluster consists of the northeast, east, and southeast regions (four out of nine counties). The posterior correlation between risks of AIDS progression and mortality post-HIV infection was shown in Table 3. The coefficient of correlation was 0.04 suggesting a weak shared geographical and overall pattern between the risks of AIDS progression and mortality post-HIV infection. Finally, with regard to modeling details, using a vague inverse Wishart,  $IW(\mathbf{R}, \nu)$  for  $\Lambda$  where  $\nu=7$  and  $\mathbf{R}=0.1\mathbf{I}_{2 \times 2}$  leads to acceptable convergence behavior.

## DISCUSSION

In this paper, we proposed a model for spatially clustered survival data in the presence of competing risks setting. The data were from Hamadan Province, Iran, from 1997 to 2011. For these data, we used a parametric proportional hazard model in a Bayesian setting which leads to a convenient interpretation of the parameters, although, the semiparametric models could also be used. The simulation study showed that the performance of our proposed model, in which two types of random effects (spatial and failure type) are incorporated simultaneously,

**TABLE 3. Posterior estimation results of model 4**

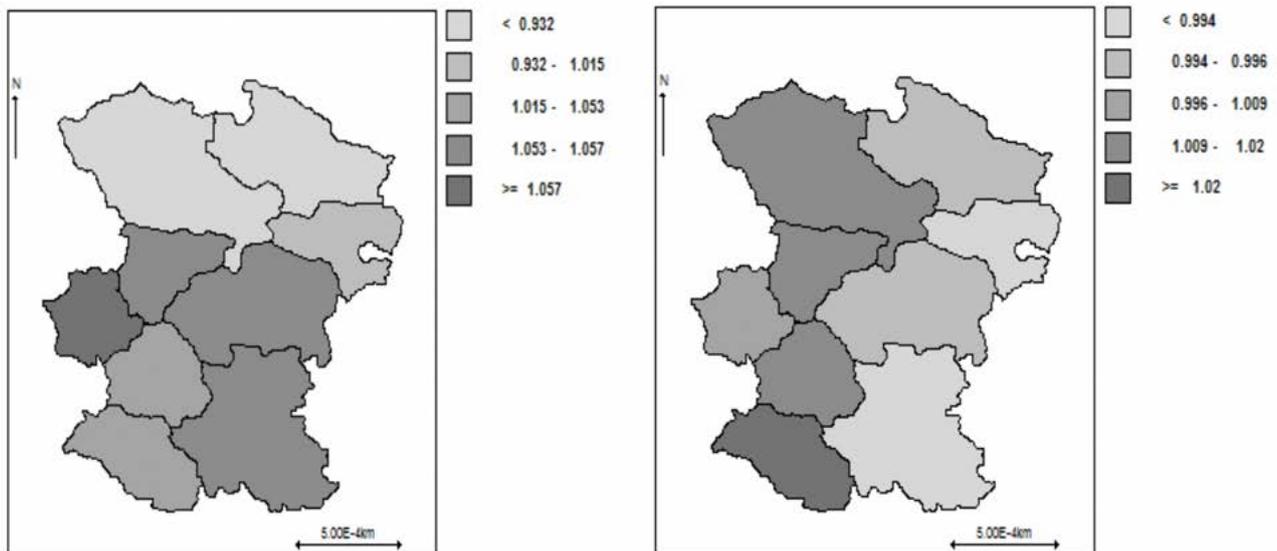
Variable	Category	Number %	AIDS progression		Mortality post-HIV infection	
			Mean (SD)	Hazard ratio (95% credible interval)	Mean (SD)	Hazard ratio (95% credible interval)
Intercept			-3.872(0.35)	0.022 (0.009, 0.039)	-3.336(0.30)	0.037 (0.018, 0.062)
Gender	Male	521 (89.05)	Reference		Reference	
	Female	64 (10.94)	0.054 (0.39)	1.14 (0.496, 2.291)	-1.759 (0.94)	0.253 (0.021, 0.844)
Marital status	Single	268 (45.81)	Reference		Reference	
	Married	223 (38.11)	0.052 (0.21)	1.079 (0.689, 1.610)	0.256 (0.20)	1.318 (0.872, 1.914)
	Divorced	75 (12.82)	0.293 (0.27)	1.394 (0.769, 2.275)	0.285 (0.28)	1.384 (0.748, 2.265)
	Widowed	19 (3.24)	0.762 (0.44)	2.357 (0.845, 4.817)	0.028 (0.64)	1.246 (0.247, 3.139)
Age	1-24	70 (11.96)	Reference		Reference	
	25-44	457 (78.11)	0.329 (0.34)	1.478 (0.741, 2.894)	0.049 (0.29)	1.099 (0.599, 1.912)
	45-74	57 (9.74)	0.824 (0.44)	2.517 (0.967, 5.554)	0.775 (0.37)	<b>2.331 (1.050, 4.555)</b>
Tuberculosis infection	No	564 (96.41)	Reference		Reference	
	Yes	21 (3.58)	1.705 (0.28)	<b>5.723 (3.077, 9.338)</b>	-0.789 (0.81)	0.601 (0.074, 1.696)
Mode of transmission	Injection drug use	475 (81.19)	Reference		Reference	
	Sexual	72 (12.30)	0.256 (0.36)	1.382 (0.614, 2.58)	-0.584 (0.56)	0.645 (0.167, 1.496)
	Mother to child	9 (1.53)	2.037(0.60)	<b>9.214 (2.296, 25.06)</b>	ND	ND
	Injection drug use/sexual	26 (4.44)	0.637 (0.36)	2.022 (0.886, 3.74)	-1.121 (0.80)	0.428 (0.052, 1.217)
$p^1=p^2$ $\Lambda_{11}$ $\Lambda_{22}$ $Corr = \Lambda_{12}/\sqrt{\Lambda_{11}\sqrt{\Lambda_{22}}}$	1.35 (0.065) 0.029 (0.035) 0.024 (0.019) 0.041 (0.038)					
ND: no data						

was good in terms of efficiency and accuracy of parameters estimation. As far as we know, our proposed model is new in survival data and could help in the visual representation of the spatial relative risk inequalities for each risk. This information also could be valuable for public health institutes and health researchers. In other words, despite the fact that we had a limited gain we analyzed the HIV/AIDS data using model 4 for its interesting clinical results. Hence, introducing this model in our dataset could be useful in similar datasets in clinical research in the future. In other words, clinical researches in the area of HIV/AIDS

can produce interesting clinical results using model 4.

According to the posterior estimates of adjusted HR, there were the significant associations between TB co-infection and mode of transmission with risk of AIDS progression. Also, there was the significant association between age at diagnosis with risk of mortality post-HIV infection. It should be mentioned that the estimated regression coefficients of cause-specific proportional hazards model give the effect of covariates on the instantaneous hazard for the type of failures and cannot be interpreted to an effect on the cumulative

**FIGURE 2. Maps of the spatial relative hazard for ADS progression (left) and mortality post-HIV infection (right) based on the model 4**



incidence function. In other words, in the cause-specific hazard model, the effect of same covariates is modeled for competing events, so there is no direct connection between the regression coefficients and the cumulative incidence functions. The reason is that the cumulative incidence function for the type of failure depends not only on the hazard of the type of failure, but also on the hazards of all other failure types. Thus, the relation of one covariate with the cumulative incidence functions of the type of failure depends on the effects of a covariate on all failure types (13, 17).

In the field of HIV/AIDS disease, understanding of geographic variation of risks of AIDS progression and mortality post-HIV infection provides greater opportunity to identify high-burden areas, since they reflect both diagnostic and patient management. From the results of present study, the low-risk cluster in risk of AIDS progression contained counties with lowest population density and the high-risk cluster in risk of AIDS progression consists of some counties with the highest rate of population density. Also, the low-risk cluster was in remote areas and with much distance from the most populous counties in Hamadan Province. In order to take more spatial variation in the HIV/AIDS data, instead of counties another smaller regions should be used, that unfortunately, were not available in our data.

Our results showed a small positive correlation between hazards of AIDS progression and mortality post-HIV infection that this information is not available when each failure type is modeled independently. We used the same variance-covariance matrix to model the MCAR and failure type random effects distributions as mentioned

before, but we could not generalize in a way that the model accounts for the different spatial and non-spatial failure type variance-covariance matrices because of the identifiability problem.

### Acknowledgments

Receiving support from the Center of Excellence in Analysis of Spatio-Temporal Correlated Data at Tarbiat Modares University is acknowledge.

### Declaration of conflicting interests

No conflict of interest has been declared by the authors with respect to the research, authorship, and/or publication of this paper.

### Funding

There is no financial support for this research.

### References

1. Motarjem K, Mohammadzadeh M, Abyar A. Geostatistical survival model with Gaussian random effect. *Statistical Papers*. 2017:1-23.
2. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*. 1979;16(3):439-54.

3. Clayton D, Cuzick J. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society Series A (General)*. 1985;82:117.
4. Banerjee S, Carlin BP, Gelfand AE. *Hierarchical modeling and analysis for spatial data*: CRC press; 2014.
5. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2002;64(4):583-639.
6. Banerjee S, Wall MM, Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*. 2003;4(1):123-42.
7. Hanson TE, Jara A, Zhao L. A Bayesian semiparametric temporally-stratified proportional hazards model with spatial frailties. *Bayesian Analysis*. 2012;7(1):147-88.
8. Banerjee S, Dey DK. Semiparametric proportional odds models for spatially correlated survival data. *Lifetime data analysis*. 2005;11(2):175-91.
9. Diva U, Dey DK, Banerjee S. Parametric models for spatially correlated survival data for individuals with multiple cancers. *Statistics in medicine*. 2008;27(12):2127-44.
10. Pan C, Cai B, Wang L, Lin X. Bayesian semiparametric model for spatially correlated interval-censored survival data. *Computational Statistics & Data Analysis*. 2014;74:198-208.
11. Cramb SM, Mengersen KL, Lambert PC, Ryan LM, Baade PD. A flexible parametric approach to examining spatial variation in relative survival. *Statistics in medicine*. 2016;35(29):5448-63.
12. Zhou H, Hanson T. A Unified Framework for Fitting Bayesian Semiparametric Models to Arbitrarily Censored Survival Data, Including Spatially Referenced Data. *Journal of the American Statistical Association*. 2018;113(522):571-81.
13. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*. 2007;26(11):2389-430.
14. Gelfand AE, Dey DK, Chang H. Model determination using predictive distributions with implementation via sampling-based methods. *STANFORD UNIV CA DEPT OF STATISTICS*; 1992.
15. Haller B, Schmidt G, Ulm K. Applying competing risks regression models: an overview. *Lifetime data analysis*. 2013;19(1):33-58.
16. Pintilie M. *Competing risks: a practical perspective*: John Wiley & Sons; 2006.
17. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*. 1999;94(446):496-509.
18. Huang X, Wolfe RA. A frailty model for informative censoring. *Biometrics*. 2002;58(3):510-20.
19. Christian NJ, Ha ID, Jeong JH. Hierarchical likelihood inference on clustered competing risks data. *Statistics in medicine*. 2016;35(2):251-67.
20. Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B (Methodological)*. 1974:192-236.
21. Besag J, Kooperberg C. On conditional and intrinsic autoregressions. *Biometrika*. 1995;82(4):733-46.
22. Torabi M. Spatial generalized linear mixed models with multivariate CAR models for areal data. *Spatial Statistics*. 2014;10:12-26.
23. Jin X, Carlin BP, Banerjee S. Generalized hierarchical multivariate CAR models for areal data. *Biometrics*. 2005;61(4):950-61.
24. Mardia K. Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*. 1988;24(2):265-84.
25. Kim H, Sun D, Tsutakawa RK. A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*. 2001;96(456):1506-21.
26. Sain SR, Cressie N. Multivariate lattice models for spatial environmental data. *Proc Statist Environ Sect Am Statist Ass*. 2002:2820-5.
27. Carlin B, Banerjee S, editors. *Hierarchical Multivariate CAR Models for Spatio-Temporal Correlated Survival Data*. *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*; 2003: Oxford University Press, USA.
28. Gelfand AE, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*. 2003;4(1):11-5.
29. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. *Statistics in medicine*. 2009;28(25):3049-67.
30. Sturtz S, Ligges U, Gelman A. *R2OpenBUGS: a package for running OpenBUGS from R*. URL <http://cran.rproject.org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS.pdf>. 2010.
31. Team RC. *R: A language and environment for statistical computing*. BM Corporation, Armonk, NY. R version 3.0.1 R Foundation for Statistical Computing, Vienna, Austria. 2015.
32. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Statistics in medicine*. 2009;28(6):956-71.
33. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical science*. 1992;7(4):457-72.