The Canadian Journal of Statistics
Vol. 40, No. 1, 2012, Pages 40–54
La revue canadienne de statistique

A Bayesian approach to mixture cure models with spatial frailties for population-based cancer relative survival data

Binbing YU[1]* and Ram C. TIWARI[2]

[1]*Laboratory of Epidemiology, Demography and Biometry, National Institute on Aging, Bethesda, MD 20892, USA*
[2]*Office of Biostatistics Center for Drug Evaluation and Research, Food and Drug Administration (FDA), Silver Spring, MD 20993, USA*

Key words and phrases: Colon cancer; cure fraction; finite mixture models; random effects; relative survival; SEER.

MSC 2010: Primary 62P10; secondary 62N02.

Abstract: As the treatments of cancer progress, a certain number of cancers are curable if diagnosed early. In population-based cancer survival studies, cure is said to occur when mortality rate of the cancer patients returns to the same level as that expected for the general cancer-free population. The estimates of cure fraction are of interest to both cancer patients and health policy makers. Mixture cure models have been widely used because the model is easy to interpret by separating the patients into two distinct groups. Usually parametric models are assumed for the latent distribution for the uncured patients. The estimation of cure fraction from the mixture cure model may be sensitive to misspecification of latent distribution. We propose a Bayesian approach to mixture cure model for population-based cancer survival data, which can be extended to county-level cancer survival data. Instead of modeling the latent distribution by a fixed parametric distribution, we use a finite mixture of the union of the lognormal, loglogistic, and Weibull distributions. The parameters are estimated using the Markov chain Monte Carlo method. Simulation study shows that the Bayesian method using a finite mixture latent distribution provides robust inference of parameter estimates. The proposed Bayesian method is applied to relative survival data for colon cancer patients from the Surveillance, Epidemiology, and End Results (SEER) Program to estimate the cure fractions. *The Canadian Journal of Statistics* 40: 40–54; 2012 © 2012 Statistical Society of Canada

Résumé: Au fur et à mesure que le traitement des cancers progresse, un certain nombre de cancers deviennent curables si le diagnostic est précoce. Dans les études de survie de population sur le cancer, une guérison se produit lorsque le taux de mortalité d'un patient cancéreux retourne à celui moyen de la population générale des individus non atteints d'un cancer. Les estimations du taux de guérison sont intéresantes pour les patients atteints d'un cancer et les gestionnaires de la santé. Les mélanges de modèles de guérison sont très utilisés, car ces modèles sont faciles à interpréter en séparant les patients en deux groupes disctincts. Habituellement, des modèles paramétriques sont utilisés pour la distribution latente des patients non guéris. L'estimation du taux de guérison à l'aide d'un mélange de modèles peut être très sensible au choix inexact de la distribution latente. Nous proposons une approche bayésienne aux mélanges de modèles de guérison pour les données de survies du cancer au niveau de la population qui peut être étendue au niveau des comtés. Au lieu d'utiliser une distribution paramétrique donnée, nous utilisons un mélange fini de distributions lognormale, loglogistique et de Weibull. Les paramètres sont estimés par la méthode de

Additional Supporting Information may be found in the online version of this article.
* *Author to whom correspondence may be addressed.*
E-mail: yubi@mail.nih.gov

Monte-Carlo markovienne. Une étude de simulation montre que la méthode bayésienne utilisant un mélange fini pour la distribution latente amène une inférence robuste des paramètres. La méthode bayésienne proposée est appliquée à des données de survie relative de patients ayant le cancer du côlon obtenues par le programme de surveillance, épidémiologie et résultats finaux (SEER) afin d'estimer les taux de guérison. *La revue canadienne de statistique* 40: 40–54; 2012 © 2012 Société statistique du Canada

1. INTRODUCTION

Cancer survival is one of the most important measures for monitoring and evaluating cancer patient care. Due to improvement in cancer treatments and dissemination of early diagnosis techniques, there has been considerable progress against cancer. For many types of cancer, a proportion of patients may be cured if diagnosed early and treated successfully. The cure fraction is defined as the proportion of patients who are cured of disease and become long-term survivors. Accounting for the cured patients may provide better predictions of long-term survival rates. Moreover, cure fraction itself is a useful measure of cancer control to researchers and policy makers.

Evaluation of cancer survival and estimation of cure fraction are often based on population-based survival data collected by cancer registries. The advantage of population-based survival analysis is that the results of such studies are representative of the entire population, a perspective which is vital for cancer control activities. However, in population-based cancer studies, cause of death may be either incorrectly identified or obtained from death certificates which are often inaccurately recorded (Begg & Schrag, 2002). For instance, it is not clear how to handle "autopsy only" cases and cases with unknown cause of death. As an alternative, relative survival (Ederer, Axtell & Cutler, 1961) is commonly used as a measure of net survival (excess mortality) due to cancer of interest. The relative survival is defined as the observed survival proportion in the patient group divided by the expected survival rate of a comparable group from the general population, who are assumed to be practically free of the cancer of interest. The expected survival can be obtained from national life table and is usually calculated after matching for age, sex, year of diagnosis, etc. The major advantage of using relative survival is that the information on the cause of death is not required, thereby circumventing problems with the inaccuracy or non-availability of death certificates. As such, relative survival has become the standard measure of estimating cancer survival for population-based cancer data.

When relative survival is used as the measure of net survival, the cure fraction is interpreted as the proportion of cancer patients whose survival experience is equivalent to the general cancer-free population. For the cancers with cure, the relative survival curve often appears to level after a number of years. This plateau implies that relative hazard is equal to one or the excess mortality rate is zero. At the point from which the cancer patients no longer experience excess mortality, we refer to the group as being "cured" (or "statistically cured"). It is important to note that this definition of cure is from a population perspective and it does not provide information on individuals. An individual may be considered "medically cured" if he or she no longer displays symptoms of the disease. However, to be certain of medical cure is difficult and in population-based cancer studies such information is unlikely to be available and thus, in the models presented here, we are only interested in cure from a population perspective.

The mixture cure model was widely used to estimate the statistical cure using the relative survival data from the population-based cancer registries. Earlier work included Boag (1949), Berkson & Gage (1952), and Cutler, Axtell & Schottenfeld (1969). Lambert et al. (2007) discussed the application of non-mixture cure models. There have been various applications of the mixture cure model. De Angelis et al. (1999) analysed the survival of Finish colon-cancer patients by adjusting for background mortality. This model was used by Mariotto et al. (2009) to estimate the number of individuals in the United States with childhood cancer. Simonetti et al. (2008) applied the mixture cure model to estimate the complete prevalence of childhood cancer.

Bejan-Angoulvant et al. (2008) discussed the advantages and limits of standard survival model and the cure model in colon cancer survival analysis. Besides the parametric mixture cure models, there are a number of semiparametric mixture cure models in the literature, for example, the proportional hazards mixture cure model (Peng & Dear, 2000; Sy & Taylor, 2000), the linear transformation cure model (Lu & Ying, 2004) and the accelerated failure time mixture cure model (Zhang & Peng, 2007, 2009; Lu, 2010). The mixture cure model has also been implemented in several statistical packages. Peng, Dear & Denham (1998) developed an R package GFCURE and Corbière & Joly (2007) provided a SAS macro for parametric and semiparametric mixture cure model. The CANSURV (Yu et al., 2004) of the National Cancer Institute (NCI) fits mixture cure models to population-based cancer survival data using likelihood-based methods.

A range of distributions including the lognormal (LN), loglogistic (LL), and Weibull (WB) distributions have been used for the latent distribution of the uncured patients. However, none of them seems to fit satisfactorily in a wide variety of samples. Finite mixture models have been used to model data with heterogeneous sub-populations. Walker & Mallick (1999) and Komarek & Lesaffre (2007) considered mixture normal distribution for cause-specific survival data and Lambert et al. (2010) used a two-component mixture Weibull distribution for population-based relative survival data. Atienza et al. (2008) used a finite mixture of different parametric families to model the length of hospital stay and proved the identifiability of such finite mixture models. One advantage of the finite mixture model based on different parametric families is the ability of accommodate different tail heaviness of sub-populations (Atienza, Garcia-Heras & Munoz-Pichardo, 2005).

In this article, we consider a Bayesian approach to the mixture cure model for grouped population-based cancer survival data. The latent distribution is modeled by a finite mixture of the union of LN, LL, and WB distributions. The parameters estimates are obtained using the Markov chain Monte Carlo (MCMC) method. One can incorporate spatial random-effects (frailties) to fit county-level cancer survival data. The estimates of county-specific cure fraction can be used for monitoring cancer survival by local policy makers. By simulation, we show that cure model with finite mixture for latent distribution can improve the estimation of cure fraction.

The rest of the article is organised as follows. In Section 2, the Bayesian mixture cure model for grouped survival data is introduced and an extension to county-level cancer survival data is developed. In Section 3, we conduct a simulation to examine the performance of the proposed Bayesian mixture cure model. The proposed method is applied to grouped cancer survival data for colon cancer patients from the SEER Program in Section 4. This article ends with a discussion in Section 5.

2. METHODS

2.1. Mixture Cure Model for Grouped Relative Survival Data

The survival function of a mixture cure model is specified as

$$S_{\mathrm{C}}(t|x) = c(x) + (1 - c(x))S_{\mathrm{U}}(t|x), \tag{1}$$

where x is the vector of covariates, $c(x)$ is the cure fraction, and $S_{\mathrm{U}}(t|x)$ is the survival function for the uncured individuals (latent distribution). A logistic model is often used to model the cure fraction

$$c(x) = \frac{\exp(x'\boldsymbol{\alpha})}{1 + \exp(x'\boldsymbol{\alpha})}, \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, .., \alpha_{L_\alpha})'$ is the vector of coefficients. The latent distribution $S_{\mathrm{U}}(t)$ could be parametric or non-parametric (piecewise constant). The commonly used parametric distributions include the LN, LL, or WB distributions and the corresponding survival functions are $S_{\mathrm{LN}}(t|x) = 1 - \Phi\big(\frac{\log t - \mu(x)}{\sigma}\big)$, $S_{\mathrm{LL}}(t|x) = \big\{1 + \exp\big[\frac{\log t - \mu(x)}{\sigma}\big]\big\}^{-1}$, and $S_{\mathrm{WB}}(t|x) = \exp\big\{-\exp\big[\frac{\log t - \mu(x)}{\sigma}\big]\big\}$. These distributions belong to a location-scale family with a location parameter μ and a scale parameter σ. The median survival time is $\exp(\mu)$ for the LN and LL distributions and is $\exp(\mu)(\log 2)^{\sigma}$ for the Weibull distribution. In general, the median of the latent survival for uncured patients is not $\exp(\mu)$ and can be obtained by setting $S(t|x) = 0.5$.

Usually, we assume that the location parameter is related to covariates as

$$\mu(x) = x'\boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{L_\beta})'$ is the vector of regression coefficients. Note that the covariates used in $c(x)$ and $\mu(x)$ may not be identical. Hence, the analysts should pick the appropriate variables to be used in each component. It is also possible to fit a non-parametric (usually piecewise) model (Peng & Dear, 2000) or a semi-parametric model (Zhang & Peng, 2007), but these are not considered here. In this article, we proposed to model the latent distribution with a three-component mixture distribution

$$S_{\mathrm{U}}(t|x) = p_1 S_{\mathrm{LN}}(t|x) + p_2 S_{\mathrm{LL}}(t|x) + p_2 S_{\mathrm{WB}}(t|x),$$

where $p_1, p_2, p_3 > 0$ and $p_1 + p_2 + p_3 = 1$. Let $\boldsymbol{\theta} = (p_1, p_2, p_3, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma)$ denote the vector of parameters to be estimated. The same parameters $\boldsymbol{\beta}$ are used in all three components of the mixture distribution $S_{\mathrm{U}}(t|x)$. By using this formulation, the effect of each covariate can be represented by a single coefficient.

When modeling relative survival, it is usually assumed that the overall hazard of dying from any cause $\lambda(t|x)$ is the sum of the expected hazard $\lambda_{\mathrm{E}}(t|x)$ and the excess hazard due to cancer $\lambda_{\mathrm{C}}(t|x)$. Under the additive hazards assumption, the overall survival can be written as a product of the expected survival, $S_{\mathrm{E}}(t|x)$, and the cause-specific survival function $S_{\mathrm{C}}(t|x)$:

$$S(t|x) = S_{\mathrm{E}}(t|x) S_{\mathrm{C}}(t|x). \tag{4}$$

For population-based cancer survival data, the additive hazards models are generally biologically more plausible and provide a better fit to the data than the multiplicative hazards models (Dickman et al., 2004).

Because of the large sample size of the population-based cancer data, the survival data are usually stratified by contiguous variables, for example, age, year of diagnosis, and cancer stages. The survival times after diagnosis are usually grouped into intervals $I_j = [t_{j-1}, t_j)$, $j = 1, 2, \ldots, J$, where t_J is the study cutoff time. Therefore, we do not have exact individual survival data, but grouped survival data with discrete survival times. Because the NCI publications usually report cancer survival rates in years. Following the convention, we choose to use annual interval in our analysis. For the patient cohort with covariates x, let n_{xj} be the number of people alive at the beginning of interval I_j. In relative survival analysis, we consider d_{xj} as the number of patients dying from all causes instead of those dying from cancer of interest and l_{xj} as the number of patients lost to follow-up in interval I_j. Using the actuarial assumption, the adjusted number of person-years at risk is $r_{xj} = n_{xj} - 0.5 l_{xj}$. The probability of dying from all causes during the interval I_j given that a subject is alive at the beginning of the interval is given by

$$\lambda_j(x) = P(T < t_j | T \geq t_{j-1}; x) = 1 - \frac{S(t_j|x)}{S(t_{j-1}|x)} = 1 - \frac{S_{\mathrm{C}}(t_j|x)}{S_{\mathrm{C}}(t_{j-1}|x)} E_{xj}, \tag{5}$$

where $S(t|x)$ is the survival function defined in (4) and $E_{xj} = \frac{S_E(t_j|x)}{S_E(t_{j-1}|x)}$ is the expected probability of surviving interval I_j that can be obtained from the life tables for the general population. The observed grouped relative survival data are denoted by $D = \{x, r_{xj}, d_{xj}, E_{xj}, j = 1, \ldots, J\}$.

A Binomial distribution is often used to model the number of deaths from all causes in each interval, that is, $d_{xj} \sim \text{Binomial}(r_{xj}, \lambda_j(x))$. The likelihood function for the grouped relative survival data is specified as

$$L(\boldsymbol{\theta}|D) = \prod_x \prod_{j=1}^{J} \lambda_j(x)^{d_{xj}}(1 - \lambda_j(x))^{r_{xj}-d_{xj}}. \tag{6}$$

Sometimes, the binomial distribution may not be an appropriate assumption for rare cancers with a smaller number of deaths or cancer sites with good survival. There are even no deaths in certain intervals. In this scenario, we advocate the use of a Poisson distribution for the observed number of deaths, that is, $d_{xj} \sim \text{Poisson}(r_{xj}\lambda_j(x))$. In this situation, the likelihood function is written as

$$L(\boldsymbol{\theta}|D) = \prod_x \prod_{j=1}^{J} (r_{xj}\lambda_j(x))^{dxj} \exp(-r_{xj} \times \lambda_j(x)). \tag{7}$$

The two approaches for the distributions of number of deaths usually give very similar results since the log-likelihoods are similar for survival data with moderate or low annual death rates (Dickman et al., 2004). In this article, we focus on grouped survival time with discrete variable x. However, the proposed method and estimation procedure can be extended to continuous survival data with continuous variables with minor modification. For example, Dickman et al. (2004) has described the estimation method for survival data with continuous variable based on likelihood (7). For individual-level survival data with continuous variable x, the number of death d_{xj} will take value 0 or 1, while the likelihood functions (6) and (7) remain the same. Furthermore, the proposed model can be used for continuous survival data with right censoring by replacing the discrete survival time j with exact survival (censoring) time t in likelihood function (7).

2.2. Extension to County-Level Cancer Survival Data

With the emergence of Geographic Information Systems, health policy makers and public health researchers started to investigate the spatial association and patterns of health outcomes. For local policy-makers, health researchers, it may also be of interest to seek detailed county-level cancer incidence data to assess the burden of cancer in local regions. There has been growing attention in detecting survival patterns and clustering by counties. Banerjee & Carlin (2003) used a hierarchical Bayesian model with spatially correlated frailties to examine infant mortality. They used standard parametric survival models without cure. Then Banerjee & Carlin (2004) extended this earlier work to spatio-temporal settings within a semiparametric model. Recently, Banerjee, Wall & Carlin (2003) developed a parametric spatial cure rate models for interval-censored survival data from smoking cessation trials and Cooner, Banerjee & McBean (2006) proposed a Bayesian modeling framework that models spatial association for geographically referenced survival data. Both methods used a non-mixture cure model, which did not consider explicitly the effect of covariates on cure fraction. Now states or local governments started to collect county-level cancer survival data and report cure fraction estimates that account for potential spatial variation. Because a covariate may have different effects on short-term and long-term survival (cure), it is ideal to model the differential effects on cure fraction and hazard rate for the uncured patients. We may capture this variation in the cure fraction or the hazard function through county-specific spatial random effects (frailties).

Suppose that there are I geographical regions and let $r_i = (r_{i1}, r_{i2})$ be the bivariate random effect (frailty), where r_{i1} denotes the frailty for the cure fraction and r_{i1} denotes the frailty for the survival for uncured patients. The cure fraction $c(x)$ and the location parameter $\mu(x)$ for the uncured patients for the ith geographic regions can be written as

$$c_i(x) = \frac{\exp(x'\boldsymbol{\alpha} + r_{i1})}{1 + \exp(x'\boldsymbol{\alpha} + r_{i1})} \tag{8}$$

and

$$\mu_i(x) = x'\boldsymbol{\beta} + r_{i2}, \tag{9}$$

respectively. One can use multivariate conditionally autoregressive (MCAR) distributions for the spatial random effects in the baseline hazard rate and the spatial frailties. Most CAR models are members of the family developed by Mardia (1988). Specifically, the MCAR of Gamerman, Moreira & Rue (2003) is specified as

$$\pi(r_1, \ldots, r_I|\Omega_r) \propto |\Omega_r|^{-I/2} \exp\left\{\sum_{i \neq j}^{I} c_{ij}(r_i - r_j)'\Omega_r^{-1}(r_i - r_j)\right\},$$

where Ω_r is the dispersion matrix and c_{ij} if areas i and j are adjacent and 0 otherwise (Congdon, 2007). Let $r_i = (r_{i1}, \ldots, r_{ip})', i = 1, \ldots, I$ be the multivariate p-dimensional ($p \geq 2$) vector of spatially correlated random effects. For mixture cure model specified in (8) and (9), $p = 2$. The intrinsic MCAR prior with 0–1 adjacency weights (Besag, York & Mollie, 1991) gives the conditional distribution

$$r_i|r_{(-i)} \sim \text{MVN}_p(\bar{r}_i, V/n_i),$$

where $r_{(-i)}$ denotes the elements of the $2 \times I$ matrix excluding the ith area (column), $\bar{r}_i = (\bar{r}_{i1}, \bar{r}_{i2})$ with $\bar{r}_{ip} = \sum_{j \in \delta_i} r_{jp}/n_i$, δ_i, and n_i denote the set of labels of the "neighbours" of area i and the number of neighbours, respectively, and V is a 2×2 dispersion matrix. The intrinsic MCAR is currently implemented in GeoBUGS, a module of WinBUGS (Lunn et al., 2000). The syntax for a MCAR distribution is

```
R[1:p,1:J] ~ mv.car(adj[], weights[], num[], omega[,])
```

where `adj[]` is a vector that represents the adjacency matrix for the study region, `weights[]` gives unnormalised weights associated with each pair of areas, `num[]` gives the number of neighbours for each area, and `omega[,]` is the precision matrix.

2.3. Bayesian Analysis

Let $\pi(\boldsymbol{\theta})$ be the joint prior distribution of $\boldsymbol{\theta}$ and let $L(\boldsymbol{\theta}|D)$ be the likelihood function for the relative survival data given by (2.2) or (2.3). The joint posterior distribution of $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}|D) \propto L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta}).$$

We assume that the prior distributions for the parameters are mutually independent, that is,

$$\pi(\boldsymbol{\theta}) = \left(\prod_{l=1}^{L_\alpha} \pi(\alpha_l)\right)\left(\prod_{l=1}^{L_\beta} \pi(\beta_l)\right)\pi(\sigma)\pi(r)\pi(p),$$

where $\pi(\alpha_l), l = 1, \ldots, L_\alpha$, and $\pi(\beta_l), l = 1, \ldots, L_\beta$, are the priors for regression coefficients, $\pi(\sigma)$ is the prior for the scale parameter, $\pi(r) = \pi(r_1, \ldots, r_I|\Omega_r)$ is the prior for the spatial random effects and $\pi(p)$ is the prior probability of being different distributions. In particular, we assume conjugate prior for the model parameters: $\pi(\alpha_j) \sim N(\alpha_{0j}, \sigma^2_{\alpha j})$, $j = 1, \ldots, L_\alpha$; $\pi(\beta_j) \sim N(\beta_{0j}, \sigma^2_{\beta j})$, $j = 1, \ldots, L_\beta$; $p(\sigma^2) \sim IG(a, b)$, where $IG(a, b)$ denotes an inverse gamma distribution with shape parameter a and scale parameter b. In addition, we also assume that the hyperprior for Ω_r is an inverse Wishart distribution with scale matrix C_γ and ν_γ degrees of freedom and assume a Dirichlet prior for the proportions $p = (p_1, p_2, p_3)$.

One can use the MCMC algorithm for parameter estimation and inferences. The proposed method can be implemented in the freely available software WinBUGS or OpenBUGS (Lunn et al., 2009). After a sufficient number of burn-in iterations, the remaining samples from the MCMC simulations are used to obtain any function of the parameters of interest. In order to see how stable the final estimates are, multiple MCMC runs are conducted with different initial values and starting points. The convergence of the MCMC samples of the parameters after excluding the initial burn-in samples are monitored using the R package CODA. For example, Gelman & Rubin (1992) used a "potential scale reduction factor" for each parameter in $\boldsymbol{\theta}$, together with upper and lower confidence limits.

3. SIMULATION

In the simulation, the overall survival function is specified as Equation (4), where the expected survival rates are extracted from US life table and the cause-specific survival function takes the form of a mixture cure model (1). To be representative of the SEER 9 registry data, which ranges from 1975 to 2002, we set the study cutoff time at 27 years.

We consider two binary covariate $x = (x_1, x_2)$, where x_1 is related to cure fraction and x_2 is related to the location parameter μ. Therefore, the covariates for the four strata are $(x_1, x_2) = (0, 0), (1, 0), (0, 1)$, and $(1, 1)$. We assume that the initial number of subjects with cancer in each stratum is $n_{x1} = 10,000$. The number of people dying in interval j is generated as a binomial variable $d_{xj} \sim \text{Binomial}(n_{xj}, \lambda_j(x))$, where $\lambda_j(x)$ is the probability of dying from all causes in Equation (5). We assume that there is no subjects lost to follow-up in each interval, therefore the number of people alive in the beginning of interval $j + 1$ is $n_{x,j+1} = n_{xj} - d_{xj}$.

For the parameters in the mixture cure model, we assume that $c(x) = \exp(\alpha_0 + \alpha_1 x_1)/[1 + \exp(\alpha_0 + \alpha_1 x_1)]$ and $\mu(x) = \beta_0 + \beta_1 x_2$. Let $c_0 = \exp(\alpha_0)/[1 + \exp(\alpha_0)]$. The parameters (c_0, e^{β_0}) are set to be (0.5, 6), (0.3, 4), and (0.1, 2) to represent cancer with good, moderate, and poor survival rates, respectively. We consider three possible underlying latent distributions, namely the LN, LL, and WB distributions described in Section 2.1. We set the parameters $(\alpha_1, \beta_1) = (1, \log(1.5)), (1, 0)$, and $(0, \log(1.5))$, representing the effects on both cure and latent distribution.

For each simulated data set, the mixture cure models with the LN, LL, and WB latent distributions are fitted and the proposed Bayesian model with finite mixture (MIX) latent distribution is also used to obtain the parameter estimates. We focus on the parameters α_1 and β_1, which measure the effect of covariates on cure fraction and survival for uncured patients, respectively. The final summary measures of the parameter estimates are based on 1,000 replications. Table 1 shows the biases of the parameter estimates and the actual coverage rate (CR) of the 95% credible intervals (CIs) for α_1 from different models. The columns for the fitted models LN, LL, and WB are the summary measures for the LN, LL, and WB model, respectively. The last two columns with header MIX correspond to the Bayesian method with mixture latent distribution.

For all cases, when the latent distribution is correctly specified, the resulting biases for α_1 are the smallest or close to the smallest. For the cases where there is a covariate effect on cure fraction ($\alpha_1 = 1$), misspecification of latent distribution leads to a remarkable bias of the parameter. In contrast, we see that the biases of α_1 based on MIX distribution are reasonably small for all cases.

TABLE 1: Biases of the parameter estimates and coverage rates of the 95% credible intervals for cure parameter α_1.

			Method							
			LN		LL		WB		MIX	
(c_0, e^{β_0})	(α_1, e^{β_1})	True model	Bias	CR	Bias	CR	Bias	CR	Bias	CR
(0.1,2)	(1,1.5)	LN	0.002	95.1	0.114	34.6	−0.185	0.0	0.027	90.9
		LL	−0.119	41.9	0.004	95.0	−0.357	0.0	−0.066	91.8
		WB	0.167	2.3	0.371	0.1	0.002	95.6	0.047	83.2
	(1,1)	LN	0.000	96.2	0.119	23.9	−0.153	0.2	0.062	82.1
		LL	−0.131	27.5	0.005	94.9	−0.331	0.0	−0.041	95.6
		WB	0.108	19.6	0.252	0.0	0.001	96.8	0.037	86.8
	(0,1.5)	LN	0.001	94.2	0.001	94.5	0.001	93.5	0.001	94.3
		LL	−0.001	95.3	−0.002	95.1	−0.001	94.9	−0.001	95.2
		WB	−0.001	94.7	−0.002	94.5	−0.001	95.0	−0.002	94.8
(0.3,4)	(1,1.5)	LN	0.001	94.7	−0.007	94.0	−0.141	0.1	−0.009	93.6
		LL	0.043	89.5	0.009	95.5	−0.187	0.0	−0.044	88.0
		WB	0.266	0.0	0.292	0.0	0.001	94.8	0.055	70.9
	(1,1)	LN	0.001	96.4	0.007	95.1	−0.122	0.2	−0.005	96.0
		LL	0.007	94.1	0.004	94.4	−0.178	0.0	−0.044	85.4
		WB	0.181	0.1	0.233	0.0	0.001	94.8	0.045	77.6
	(0,1.5)	LN	−0.000	94.4	−0.000	94.7	−0.000	94.7	−0.000	94.5
		LL	0.000	94.8	0.000	94.9	0.000	94.7	0.000	94.6
		WB	0.000	95.2	−0.000	95.0	−0.000	94.9	−0.004	95.1
(0.5,6)	(1,1.5)	LN	−0.026	84.1	−0.023	90.4	−0.161	28.4	−0.021	88.0
		LL	0.087	72.8	0.004	94.8	−0.089	40.6	−0.030	92.9
		WB	0.314	0.0	0.192	2.1	0.002	94.2	0.064	75.5
	(1,1)	LN	0.000	95.4	−0.020	92.6	−0.079	28.6	−0.022	93.6
		LL	0.054	83.2	0.004	95.1	−0.101	27.0	−0.030	90.4
		WB	0.214	0.6	0.165	2.1	0.001	94.3	0.052	79.0
	(0,1.5)	LN	−0.044	94.2	0.002	95.0	−0.018	94.3	−0.001	95.0
		LL	0.000	95.6	−0.000	95.6	0.000	95.9	0.000	95.5
		WB	0.000	95.1	−0.000	95.0	0.000	95.3	0.000	95.4

For example, in the first row where the true latent distribution is LN, the bias from the correctly specified LN model is 0.002 and the biases of α_1 are 0.114 and −0.185 when the fitted models are LL and WB, respectively. While the bias for MIX is only 0.027. This shows that the cure method with mixture latent distribution is a robust approach for estimating cure fraction. When the latent distribution is correctly specified, the actual CR of the 95% CI is close to the nominal level. However, when the latent distribution is misspecified, the actual CR rate may be far below the 95% level. This is largely due the biases of the parameter estimates from the misspecified models. Compared to the other misspecified models, the CIs from the cure model with mixture latent distribution maintain a CR closer to the 95% level. When the covariate has no effect on cure fraction ($\alpha_1 = 0$), the estimates from all models are empirically unbiased and the 95% CIs have CR close to the nominal level. From the frequentist point of view, this indicates that the test of covariate effect on cure fraction maintain a significance level close to 0.05 regardless whether the latent distribution is misspecified.

We also compared the empirical standard deviation of the posterior estimates and the mean of the estimated standard errors from the posterior samples for α_1, we found that the two estimates of standard deviations for all cases are rather close (see Table A in Supplementary Document). In addition, we evaluated the biases and coverage rates of the 95% CIs from different models for parameter β_1. Compared to the cure parameter α_1, the biases for β_1 from different models are smaller (see Tables B and C in Supplementary Document). However, we still see similar pattern that misspecified models may yield a larger bias while the cure model with mixture latent distribution is more robust with smaller bias and CR closer to the nominal level.

Moreover, we conducted simulations to examine the performance of the proposed estimation method for county-level cancer data with spatial frailties. The simulation was based on the 99 counties in the state of Iowa. Let $D = \text{Diag}(m_i)$ be the diagonal matrix of the number of neighbours m_i for each area and let W be the adjacency matrix based on the Iowa county map. We assume that vector of frailties

$$r_j \overset{\text{iid}}{\sim} N(\mathbf{0}, [D - \xi_j W]^{-1}), \; j = 1, 2,$$

where $r_j = (r_{1j}, \dots, r_{Ij})$ are the spatial frailties specified in (8) and (9), $\xi_j = 0.5$ is the scale parameter to ensure the propriety of the CAR distributions. The bias and the CR of the 95% CIs for the cure parameter α_1 are shown in Table 2. Again, we see that the misspecification of the latent distribution may yield a large amount of bias and lead to gross under-coverage of the 95% CI. The cure method with mixture latent distribution remains robust with respect to different latent distributions and has CR closer to the nominal level then the other models.

In the simulations described above, the discrete survival data were generated based on models (4) and (5). To examine the effect of grouping continuous survival times, we also simulated continuous survival data and applied the mixture cure models with properly grouped survival times. First, survival data were generated from two competing risks, one was from the expected survival for cancer-free population, the other was from the mixture cure model from the cancer-specific survival function (1). Then the continuous survival times were grouped into annual intervals with maximum follow-up time 27 years. The simulation results remained similar (see Tables D and E in Supplementary Document). Overall, the simulations showed that if the latent distribution in the mixture cure model was misspecified, the resulting parameters, especially the

TABLE 2: Bias of the cure parameter and the coverage rate of the 95% credible intervals.

	Fitted model							
	LN		LL		WB		MIX	
True model	Bias	CR	Bias	CR	Bias	CR	Bias	CR
LN	−0.008	91.0	0.009	62.2	−0.062	5.8	−0.003	84.3
LL	−0.683	57.4	−0.002	93.0	−0.113	3.1	−0.047	92.7
WB	0.073	1.6	0.118	1.1	−0.024	95.2	−0.032	87.6

parameters for cure fraction, could have significant amount of bias. The parameter estimates based on cure model with mixture latent distribution had much less bias.

4. APPILICATIONS

The Surveillance, Epidemiology, and End Results (SEER) Program of the NCI is an authoritative source of information on cancer incidence and survival in the U.S. Starting from 1973 to 1975, the SEER Program included nine registries covering almost 10% of the US population (SEER 9 registries), and then went through several expansions in the early 1990s and 2000 (SEER 13 and SEER 17 registries), respectively. Currently, the SEER 17 registries, excluding Alaska, cover about 26% of the total US population. Because of confidentiality issues, the SEER Program does not disclose the exact survival time for each subject, but the survival months. The SEER Program provides the number of patients alive at the beginning of a time period (month or year), the observed number of deaths during this period, and the cause of their death.

4.1. Relative Survival for Colon Cancer Patients

We apply the mixture cure model to estimate the survival and cure fraction for colon cancer patients obtained from the SEER program. To avoid the inaccurate specification of cause of death, in the following, we conduct only the relative survival analysis. We use the patients diagnosed with colon cancer between 1988 and 2003. The maximum follow-up time is 15 years. The relative survival data are stratified by race (white vs. black), sex, American Joint Committee on Cancer (AJCC) stage, and grade. The AJCC stage is the most commonly used staging system for colorectal cancer, sometimes also known as the TNM system. The stage is expressed in Roman numerals from stage I (the least advanced) to stage IV (the most advanced). Another factor that can affect the outlook for survival is the grade of the cancer. Grade is a description of how closely the cancer resembles normal colorectal tissue when looked at under a microscope. The scale used for grading colorectal cancers goes from G1 (where the cancer looks much like normal colorectal tissue) to G4 (where the cancer looks very abnormal). The grades G2 and G3 fall somewhere in between. The grade is often simplified as either "low-grade" (G1 or G2) or "high-grade" (G3 or G4). Low-grade cancers tend to grow and spread more slowly than high-grade cancers. Most of the time, the outlook is better for low-grade cancers than it is for high-grade cancers of the same stage. Doctors sometimes use this distinction to help decide whether a patient should get additional (adjuvant) treatment with chemotherapy after surgery.

In the analysis, the total number of patients diagnosed with colon cancer is 90,398. The four variables used for stratification are used as covariates in both cure fraction and latent distribution as specified in (2) and (3). The reference group are the white male patients diagnosed with stage I and grade 1 colon cancer. The posterior probabilities of the latent distribution being LN, LL, and WB are (0.125, 0.125, and 0.75). The posterior mean, standard errors (SE), and 95% CIs of the parameters are shown in Table 3. For the reference group, the cure fraction estimate is 95.7% (95% CI: 80.7–99.4%) and the median survival time for the uncured is 16.3 years (95% CI: 3.1–30.0). Compared to the white patients, the blacks have significantly lower cure fractions, but insignificantly shorter survival time for the uncured patients. Compared to the male, the female patients have significantly lower cure fraction, but the survival times for the uncured patients are similar for both sexes. This shows the survival disparity in different races and sexes still exist after adjusting for AJCC cancer stage and grade. Compared to stage I cancer patients, the patients with stage III and IV cancers have significantly lower cure fractions and shorter survival time for the uncured. The difference of cure fraction and survival time for the uncured between grade 1 and 2 are not significant. But the patients with grade 3–4 have much poor survival in terms of both cure fraction and survival time if uncured.

TABLE 3: Parameter estimates for the colon cancer data.

	Latent distribution			Cure		
Variable	Mean	SE	(95% CI)	Mean	SE	(95% CI)
Intercept	3.095	1.336	(1.429, 5.114)	0.443	2.206	(−3.848, 2.379)
Race						
White						
Black	−0.033	0.080	(−0.195, 0.149)	−0.430	0.234	(−1.215, −0.061)
Sex						
Male						
Female	−0.151	0.072	(−0.319, −0.067)	0.801	1.276	(0.028, 3.389)
SEER AJCC stage						
I						
II	−0.812	0.517	(−1.451, −0.005)	−0.673	0.746	(−1.181, −0.197)
III	−1.403	0.996	(−2.376, −0.028)	−1.379	0.721	(−2.414, −0.631)
IV	−2.691	1.305	(−4.589, −1.059)	-4.469	0.861	(−5.72, −3.275)
Grade						
G1						
G2	−0.101	0.058	(−0.204, −0.002)	−0.208	0.176	(−0.693, −0.008)
G3–G4	−0.641	0.088	(−0.815, −0.482)	−0.368	0.112	(−0.547, − 0.207)
Scale parameter	1.082	0.116	(0.964, 1.302)			

TABLE 4: Mean, minimum, and maximum cure fraction and median survival time for the uncured patients.

	Cure fraction (%)			Median survival for uncured		
	Minimum	Mean	Maximum	Minimum	Mean	Maximum
Localised	81.9	85.4	87.4	0.7	1.9	4.3
Regional	47.6	54.1	58.3	0.4	1.1	2.5
Distant	3.4	4.4	5.2	0.2	0.7	1.5

4.2. Estimation of County-Specific Cure Fraction of Colon Cancer in Iowa

We estimated the county-specific cure fraction for colon cancer in Iowa. The data were extracted from SEER*Stat software. For each of the 99 counties, the survival data were stratified by categorical historical stage (x), that is, localised, regional, and distant stages. The SEER historical stage was used for county-level data because the information on historical stage was collected since 1973, which allowed enough number of patients in each county for accurately estimating the parameters. We assume that both $c_i(x)$ and $\mu_i(x)$ are related to historic stage. The posterior probabilities of the latent distribution being LN, LL, and WB are 0.52, 0.14, and 0.34, respectively.

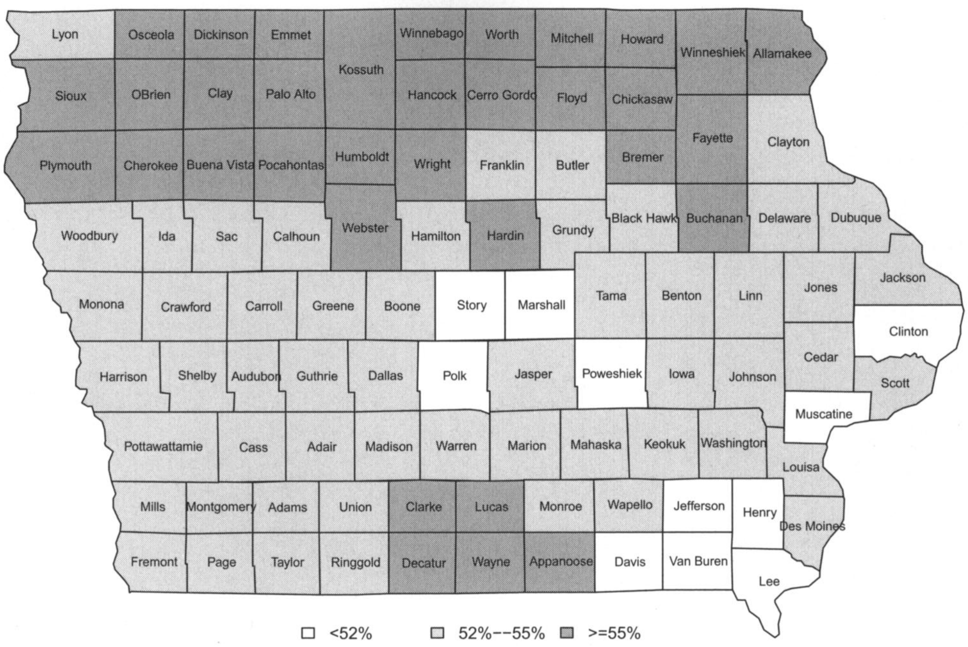

FIGURE 1: Map of cure fraction for regional colon cancer in Iowa.

Table 4 shows the mean, minimum, and maximum of the cure fractions and median survival times for the uncured patients for the 99 counties by SEER historic stage.

For localised colon cancer, the cure fractions for different counties are rather homogeneous with mean 85.4% and range (81.9%, 87.4%). The mean and range of the cure fractions for regional colon cancer are 54.1% and (47.6%, 58.3%). For distant colon cancer, the mean cure fraction is 4.4% with range (3.4%, 5.2%). There is slightly more variation of the cure fractions for the regional colon cancer. Therefore, we show the map of the county-level cure fraction for regional colon cancer in Figure 1. The cure fractions are grouped into three categories. The white, light grey, and dark grey represents the groups with cure fractions <52%, 52–55%, and >55%, respectively. From this map, we can see that the northern Iowa counties have relatively higher cure fractions than the southern counties. This may provide useful information to health care planners for allocating limited resources for cancer control.

5. DISCUSSION

We propose a Bayesian MCMC method for fitting mixture cure model for grouped relative survival data from the population-based cancer registries. The simulation study shows the robustness of using finite mixture distribution for latent distribution for parameter estimation and inference. The proposed method is easy to implement and can be used for county-level survival data. By using spatial random effects, it allows us to borrow the information from across all counties in estimation of cure fraction and survival estimates.

In the cure model, a three-component mixture with LN, LL, and WB models is chosen for the latent survival distribution. There are two main reasons. First, the three distributions are the most commonly used distributions in survival analysis. Second, the mixture of different components is used to increase the flexibility of accommodating distributions with different shapes and tails. The third reason is because of the computation convenience. Theoretically, a two or four component mixture can also be used and the distributions can take other forms. The selection of distributions and the determination of optimal number of components require further research.

As geographically adjacent areas (counties) usually share similar environmental and social factors, spatial dependence, and clustering exist for nearby neighbourhoods. Therefore, it is natural to incorporate the spatial correlation using frailties in survival analysis. The frailties account for excess heterogeneity as well as the similarity in the cure fraction and latent survival. As individual-level or point-referenced data have become increasingly common in public health studies with the use of the Geographic Information System (GIS) technology and geocoding of individual addresses. Now county-level characteristics, for example, race-ethnicity composition and median household income, are available in SEER data, these factors are strongly related to cancer incidence and survival for a certain number of cancers. Therefore, from the public health perspective, it is useful to use the geographical information to estimate county-level survival and cure rates.

The proposed method is developed specifically for relative survival data, where the cause of death is not available. The usual competing-risk data is based on the assumption that the cause of death is known. We expect that the results would be similar if cause-specific competing risk data with known risk types were used. The relative survival is the preferred measure of cancer survival when the cause of death is not reliable. However, the cancer patients may have higher risk of dying from other causes than the general population. When this assumption is true, then the relative survival (or excess mortality) cannot be interpreted as cancer-specific survival. This is a disadvantage of relative survival analysis when the goal is to estimate the cause-specific sub-survival function.

ACKNOWLEDGEMENTS

The authors wish to thank two anonymous reviewers and the Associate Editor for their helpful comments. The views expressed in this article by the author, Dr. Tiwari, are solely his own and do necessarily reflect that of the FDA. Dr. Yu was supported in part by the Intramural Research Program of the National Institute on Aging, National Institutes of Health. This research utilised the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, Maryland, USA (http://biowulf.nih.gov).

BIBLIOGRAPHY

Atienza, N., Garcia-Heras, J., & Munoz-Pichardo, J. M. (2005). A new condition for identifiability of finite mixture distributions. *Metrika*, 63, 215–221.

Atienza, N., Garcia-Heras, J., Munoz-Pichardo, J. M., & Villa, R. (2008). An application of mixture distributions in modelization of length of hosptical stay. *Statistics in Medicine*, 27, 1403–1410.

Banerjee, S. & Carlin, B. P. (2003). Semiparametric spatiotemporal frailty modelling. *Environmetrics*, 14, 523–535.

Banerjee, S. & Carlin, B. P. (2004). Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics*, 60, 268–275.

Banerjee, S., Wall, M., & Carlin B. P. (2003). Frailty modelling for spatially correlated survival data with application to infant mortality in Minnesota. *Biostatistics*, 4, 123–142.

Begg, C. B. & Schrag D. (2002). Attribution of deaths following cancer treatment. *Journal of National Cancer Institute*, 94(14), 1044–1045.

Bejan-Angoulvant, T., Bouvier, A. M., Bossard, N., Belot, A., Jooste, V., Launoy, G., & Remontet, L. (2008). Hazard regression model and cure rate model in colon cancer relative survival trends: are they telling the same story? *European Journal of Epidemiology*, 23, 251–259.

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47, 501–515.

Besag, J., York, J., & Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B*, 11, 15–44.

Congdon, P. (2007). Bayesian modelling strategies for spatially varying regression coefficients: A multivariate perspective for multiple outcomes. *Computational Statistics and Data Analysis*, 51(5), 2586–2601.

Cooner, F., Banerjee, S., & McBean, A. M. (2006). Modelling geographically referenced survival data with a cure fraction. *Statistical Methods in Medical Research*, 15(4), 307–324.

Corbière, F. & Joly, P. (2007). A SAS macro for parametric and semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 85(2), 173–180.

Cutler, S. J., Axtell, L. M., & Schottenfeld, D. (1969). Adjustment of long-term survival rates for deaths due to intercurrent disease. *Journal of Chronic Diseases*, 22(6–7), 485–491.

De Angelis, R., Capocaccia, R., Hakulinen, T., Soderman, B., & Verdecchia, A. (1999). Mixture models for cancer survival analysis: Application to population-based data with covariates. *Statistics in Medicine*, 18(4), 441–454.

Dickman, P., Sloggett, A., Hills, M., & Hakulinen T. (2004). Regression models for relative survival. *Statistics in Medicine*, 23(1), 51–64.

Ederer, F., Axtell, L. M., & Cutler, S. J. (1961). The relative survival rate: A statistical methodology. *National Cancer Institute Monograph*, 6, 101–121.

Gamerman, D., Moreira, A. R. B., & Rue, H. (2003). Space-varying regression models: Specifications and simulation. *Computational Statistics and Data Analysis*, 42(3), 513–533.

Gelman, A. & Rubin, D. (1992). Inference from alternative simulation using multiple sequences. *Statistical Science*, 7, 457–472.

Komarek, A. & Lesaffre, E. (2007). Bayesian accelerated failure time model for correlated censored data with a normal mixture as an error distribution. *Statistica Sinica*, 17, 549–569.

Lambert, P. C., Dickman, P. W., Weston, C. L., & Thompson, J. R. (2010). Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 59(1), 35–55.

Lambert, P. C., Thompson, J., Weston, C. L., & Dickman, P. W. (2007). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8(3), 576–594.

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica*, 20, 661–674.

Lu, W. & Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, 91, 331–343.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

Lunn, D. J., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, 28, 3049–3082.

Mardia, K. V. (1988). Multi-dimensional multivariate gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24, 265—284.

Mariotto, A. B., Rowland, J. H., Yabroff, K. R., Scoppa, S., Hachey, M., Ries, L., & Feuer, E. J. (2009). Long-term survivors of childhood cancers in the United States. *Cancer Epidemiology Biomarkers and Prevention*, 18(4), 1033–1040.

Peng, Y. & Dear, K. B. G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1), 237–243.

Peng, Y., Dear, K. B. G., & Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, 17, 813—830.

Simonetti, A., Gigli, A., Capocacia, R., & Mariotto, A. (2008). Estimating complete prevalence of cancers diagnosed in childhood. *Statistics in Medicine*, 27(7), 990–1007.

Sy, J. P. & Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56, 227–236.

Walker, S. & Mallick, B. K. (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics*, 55, 477–483.

Yu, B., Tiwari, R. C., Cronin, K. A., & Feuer, E. J. (2004). Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine*, 23(11), 1733–1747.

Zhang, J. & Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, 26(16), 3157–3171.

Zhang, J. & Peng, Y. (2009). Accelerated hazards mixture cure model. *Lifetime Data Analysis*, 15(4), 455–467.

Received 30 November 2010
Accepted 30 August 2011