Published in final edited form as:

Biom J. 2016 September ; 58(5): 1164–1177. doi:10.1002/bimj.201500111.

Semiparametric Bayesian estimation of quantile function for breast cancer survival data with cured fraction

Cherry Gupta^{*,1}, Juliana Cobre², Adriano Polpo³, Debjayoti Sinha¹

¹Department of Statistics, Florida State University, Tallahassee, FL 32306, U S A

²Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Av. Trabalhador São-carlense, São Carlos 13566-590, Brazil

³Department of Statistics, Federal University of São Carlos, São Paulo 13565-905, Brazil

Abstract

Existing cure-rate survival models are generally not convenient for modeling and estimating the survival quantiles of a patient with specified covariate values. This paper proposes a novel class of cure-rate model, the transform-both-sides cure-rate model (TBSCRM), that can be used to make inferences about both the cure-rate and the survival quantiles. We develop the Bayesian inference about the covariate effects on the cure-rate as well as on the survival quantiles via Markov Chain Monte Carlo (MCMC) tools. We also show that the TBSCRM-based Bayesian method outperforms existing cure-rate models based methods in our simulation studies and in application to the breast cancer survival data from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) database.

Keywords

Generalized Box-Cox; Markov Chain Monte Carlo; Transform both sides

1 Introduction

In various oncology trials, previous works have introduced a variety of cure-rate survival models that can accommodate a “cured fraction” (often called “cure-rate”) to account for patients that may never experience the event of interest. The earliest cure-rate survival model was the mixture cure-rate model (MCRM) originally proposed by Boag (1949) and by Berkson and Gage (1952). Based on a biological mechanism of tumor metastasis, some researchers proposed the bounded cumulative hazard model (BCHM) (Yakovlev *et al.*, 1993; Tsodikov, 1998; Ibrahim *et al.*, 2001). More recent papers such as Yin and Ibramin (2005) and Fonseca *et al.* (2013) further developed the BCHM. Some authors like Cooner *et al.* (2007) also unified MCRMs and BCHMs using the latent activation model (LACRM),

*Corresponding author: cgupta@stat.fsu.edu, Phone: +00-850-644-3518, Fax: +00-850-644-5271.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

Conflict of interest

The authors have declared no conflict of interest.

which assumes that each patient has an unknown unobservable patient-specific number of “latent factors”, that may lead to the patient’s event over time.

However, these latent factors-based models have several drawbacks. The distribution of these latent factors cannot be identified from the observed survival data. This creates difficulties in validating the model’s general structure and parameter estimates. The indirect relationship between the model’s underlying stochastic mechanism and the observable survival time makes it also difficult to interpret the effects of patients’ covariates on the observable quantities, especially when the underlying latent activation scheme is suspect for certain cancer studies.

Whether using either MCRM or BCHM, cure-rate survival models generally focus on scale and hazard ratio parameter estimation and estimating the probability of cure, but fail to address the estimation of the quantile function for any patient with arbitrarily chosen covariate values. Patients and physicians are often interested in this quantile function, because it allows the clinicians to give a quick evaluation and prediction of a patient’s survival time, rather than just the estimated probability of cure.

Frequentist quantile regression methods for survival (e.g. Portnoy, 2003; Peng and Huang, 2008) can estimate any particular linear quantile regression function of interest, however, these methods do not accommodate survival models with cured fraction. Wu and Yin (2013) extended the linear quantile regression method of Peng and Huang (2008) to accommodate survival data with cured fraction. They used bootstrap tools for variance estimation. As an alternative to these, Bayesian methods allow integration of prior information and are nowadays relatively simple to implement using publicly available software for Markov chain Monte Carlo (MCMC) tools. Bayesian methods also allow inference based on finite samples without requiring the asymptotic justifications. Bayesian quantile regression methods have been implemented and developed in various areas of statistical analysis (Kottas and Gelfand, 2001; Kottas and Krnjajic, 2009; Reich *et al.*, 2010; Todkar and Kadane, 2011; Lin *et al.*, 2012; Reich and Smith, 2013) including survival analysis. However, existing Bayesian quantile regression methods do not accommodate survival models with cure-rate.

In this paper, we propose a new cure-rate survival model which we call the transform-both-sides cure-rate model (TBSCRM) because it reduces to the transform-both-sides (TBS) survival model of Lin *et al.* (2012) when the cured fraction is 0. Therefore, our TBSCRM has all the advantages of the TBS model, including, for example, the ability to accommodate highly skewed survival data (particularly important in presence of cured fraction). Both models can incorporate a wide class of shapes for the survival function. Our model can also handle covariate effects on scale and shape without requiring any latent risk assumptions. This allows straightforward practical interpretation of the regression effects on survival, as well as direct interpretation of the covariate effects on the cure-rate. Compared to the method in Wu and Yin (2013), our method provides expressions and simultaneous estimates of all the quantile functions, permitting a coherent interpretation of the covariate effects on all quantiles of survival time. Unlike the method of Wu and Yin (2013) aiming to fit a log-linear regression line to any particular quantile of an uncured patient’s survival time, the TBSCRM presents a full survival model with the transformation parameter λ and covariate

effects parameter β . For the special case of $\lambda = 1$, TBSCRM produces the model with log-linear regression function for all quantiles of an uncured patient.

We develop a semiparametric Bayesian method for inference using TBSCRM for cure-rate survival data, and we use a Dirichlet scale-mixture of Gaussian kernel as a prior model for the nonparametric part of our model. Dirichlet process and related Polya trees are very popular priors used in nonparametric and semiparametric Bayesian inferences (e.g. Hanson and Johnson, 2002; Hjort and Walker, 2009). For survival model without cure-rate, the semiparametric Bayesian median regression method of Kottas and Gelfand (2001) showed that the Dirichlet process mixture can allow for skewness and general tail behavior.

The rest of this paper is organized as follows. In Section 2, we demonstrate both the form of the TBSCRM and its quantile function. Section 3 adds additional theoretical details of the TBSCRM such as its likelihood function and methods for setting prior distributions for the model parameters. In Section 4, we demonstrate the semiparametric Bayesian estimation of the TBSCRM and its ability to estimate the true parameters of simulated datasets with acceptable error rate, bias, and variation of the estimates. Section 5 illustrates the application of our methods to breast cancer data from the National Cancer Institute. For this application, we demonstrate the estimation of the quantile function and the better fit of the Bayesian TBSCRM to the data, as compared to other existing cure-rate survival models.

2 Model description

2.1 Distribution

Let T_i be the nonnegative random variable denoting the survival time of patient i with covariate vector $\mathbf{x}_i$. We assume that the parameter vector $\boldsymbol{\gamma}$ models the relationship between $\mathbf{x}_i$ and the probability of being cured as $\eta(\mathbf{x}_i; \boldsymbol{\gamma}) = P[T_i = +\infty | \mathbf{x}_i] = 1 / \{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\gamma})\}$. For the sake of brevity, we will often use the notation $\eta(\mathbf{x})$ instead of $\eta(\mathbf{x}; \boldsymbol{\gamma})$ even though it is actually a parametric function of covariate $\mathbf{x}$ with unknown parameter $\boldsymbol{\gamma}$. Our transform-both-sides cure-rate model (TBSCRM) of T_i is

$$P[T_i > t | \mathbf{x}_i] = S(t | \mathbf{x}_i) = \eta(\mathbf{x}_i) + \{1 - \eta(\mathbf{x}_i)\} S_*(t | \mathbf{x}_i), \quad (1)$$

where the proper survival function $S_*(t | \mathbf{x})$ of a noncured patient has the form

$$P[T_i > t | \mathbf{x}_i; T_i < +\infty] = S_*(t | \mathbf{x}_i) = S_\varepsilon \{g_\lambda(\log(t)) - g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta})\}, \quad (2)$$

of transform-both-sides (TBS) model of Lin *et al.* (2012), with $S_\varepsilon(u) = P[\varepsilon > u]$ corresponds to a unimodal and symmetric density (PDF) $f_\varepsilon(\cdot)$ centered at 0, with corresponding cumulative distribution function (CDF) $F_\varepsilon(u) = 1 - S_\varepsilon(u)$. In (2), $g_\lambda(y)$ is the monotone power transformation function (Bickel and Doksum, 1981),

$$g_\lambda(y) = \frac{y|y|^{\lambda-1}}{\lambda}, \quad (3)$$

where the transformation parameter λ is an unknown positive real number. As a consequence of (2), $S_*(t | \mathbf{x})$ in (1) is a proper survival function because $S_*(+\infty | \mathbf{x}) = 0$, and $\boldsymbol{\beta}$

models the relationship between covariate-vector* $\mathbf{x}_i$ and the survival time of an uncured patient. For the sake of the brevity of the presentations, we now focus on the case where the same covariate vector $\mathbf{x}_i$ affects both the cured fraction $\eta(\mathbf{x}_i)$ (via $\boldsymbol{\gamma}$) and survival time of an uncured patient (via $\boldsymbol{\beta}$), however, extensions of TBSCRM of (1) to the general case with two different set of covariates for cure-rate and S^* are straightforward. Lin *et al.* (2012) pointed out that $g_\lambda(y)$ is a very ideal transformation to transform a highly skewed and heteroscedastic random variable to a random variable with symmetric and unimodal density with common variance.

2.2 Quantile function

If the patient i is known to be uncured (i.e., $T_i < \infty$ is known), then the conditional survival function of this patient is (2), and following Lin *et al.* (2012) we can show that the median of this conditional survival time is $M_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$. As mentioned earlier, a major advantage of TBSCRM of (1) is that all quantiles of the survival time has the expression

$$Q_*(\tau | \mathbf{x}_i) = \exp\{g_\lambda^{-1}(g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta}) + \varepsilon_\tau^*)\}, \quad (4)$$

where ε_τ^* is the τ -quantile of the density $f_\varepsilon(\cdot)$, such that $P[\varepsilon < \varepsilon_\tau^*] = \tau$. A derivation of (4) is in the Appendix. The conditional quantile-function of (4) ensures that $P(T_i < Q_*(\tau | \mathbf{x}_i) | T_i < +\infty; \mathbf{x}_i) = \tau$. This conditional quantile function is related to the quantile function $Q(\tau | \mathbf{x}_i)$ of unconditional survival time T_i via

$$P[T_i < Q_*(\tau | \mathbf{x}_i) | \mathbf{x}_i] = \tau \{1 - \eta(\mathbf{x}_i)\} \Rightarrow Q_*(\tau | \mathbf{x}_i) = Q(\{1 - \eta(\mathbf{x}_i)\} \tau | \mathbf{x}_i) \quad (5)$$

for $0 < \tau < 1$. This implies that $\exp(\mathbf{x}_i^T \boldsymbol{\beta})$ is the $0.5\{1 - \eta(\mathbf{x}_i)\}$ -th quantile of the unconditional survival time T_i . For example, when the cure-rate is $\eta(\mathbf{x}_i) = 0.4$, then $\exp(\mathbf{x}_i^T \boldsymbol{\beta})$ is the 0.3-th quantile of the unconditional survival time. For the special case when $\eta(\mathbf{x}_i) = 0$ (no cure-rate), the TBSCRM of (1) is the same as the TBS model of Lin *et al.* (2012). When $\lambda = 1$, all quantile functionals of conditional log-survival time (as a function of covariates $\mathbf{x}_i$) of TBSCRM are linear and parallel to each other. Unlike existing quantile regression methods including that of Wu and Yin (2013), the TBSCRM of (1) allows only the median functional $Q_*(0.5 | \mathbf{x}_i)$ of an uncured patient's survival time to be log-linear and rest of the conditional and all of the unconditional quantile functions are non-linear. This is more realistic in practice than assuming all quantiles of $\log(T_i)$ to be linear functionals of $\mathbf{x}_i$. It also allows simultaneous estimation of all the quantile functionals using (4).

3 Likelihood and estimation

3.1 Likelihood

The vector of observations is $(t_i, \delta_i, \mathbf{x}_i)$ for patient $i = 1, 2, \dots, n$, where $t_i = \min(T_i, C_i)$ is the observed survival time, $\delta_i = I(T_i \leq C_i)$ is the failure indicator, and C_i is the noninformative censoring time. The likelihood function for TBSCRM of (1) is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, F_\epsilon | \boldsymbol{\delta}, \mathbf{y}, \mathbf{x}) = \prod_{i=1}^n \left[\left\{ 1 - \eta(\mathbf{x}_i; \boldsymbol{\gamma}) \right\}^{\delta_i} \left\{ f_\epsilon(g_\lambda(y_i) - g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta})) |y_i|^\lambda - 1 \right\}^{\delta_i} \times \right. \\ \left. \times \left\{ 1 - (1 - \eta(\mathbf{x}_i; \boldsymbol{\gamma})) F_\epsilon(g_\lambda(y_i) - g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta})) \right\}^{1 - \delta_i} \right], \quad (6)$$

where $y_i = \log(t_i)$, $f_\epsilon(\cdot)$ and $F_\epsilon(\cdot)$ are given in (2), $\eta(\mathbf{x}_i; \boldsymbol{\gamma})$ is the probability of being cured, and $g_\lambda(\cdot)$ is described in (3). The corresponding posterior distribution is given as

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, F_\epsilon | \boldsymbol{\delta}, \mathbf{y}, \mathbf{x}) \propto L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, F_\epsilon | \boldsymbol{\delta}, \mathbf{y}, \mathbf{x}) \pi_1(\boldsymbol{\beta}) \pi_2(\boldsymbol{\gamma}) \pi_3(\lambda) \pi_4(F_\epsilon), \quad (7)$$

where π_1 , π_2 , π_3 , π_4 are assumed independent prior distributions of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, λ , and $F_\epsilon(\cdot)$, respectively. Analytical integration of $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, F_\epsilon | \boldsymbol{\delta}, \mathbf{y}, \mathbf{x})$ is intractable and all posterior inferences are performed using MCMC samples (Gelfand and Smith, 1990) from the joint posterior in (7). For the Bayesian analysis in this paper, we implement the MCMC algorithm using publicly available software JAGS (Plummer, 2003).

3.2 Prior distributions

For our semiparametric Bayesian inference using posterior in (7), we use independent parametric prior densities for $\pi_1(\boldsymbol{\beta})$, $\pi_2(\boldsymbol{\gamma})$, and $\pi_3(\lambda)$, and a prior process $\pi_4(F_\epsilon)$ for nonparametric F_ϵ . We model the unimodal and symmetric f_ϵ as the Dirichlet scale-mixture of Gaussian kernel with

$$f_\epsilon(u) = \int_0^\infty \phi(u|\sigma) dG(\sigma), \quad (8)$$

where $G(\sigma)$ is some unknown mixing distribution and $\phi(\cdot|\sigma)$ is a Gaussian density with mean 0 and standard deviation σ . To obtain a prior process $\pi_4(F_\epsilon)$ on F_ϵ of (8), we use the Dirichlet process (DP) prior $G \sim DP(G_0, \alpha)$ (Ferguson, 1973) as the prior for the unknown $G(\sigma)$, where G_0 is the prior mean (“prior guess”) of G and α is the precision of the prior guess. A comprehensive discussion of how to select an appropriate G_0 based on a desired “prior guess” for F_ϵ is in Lin *et al.* (2012). We use $\alpha = 1$ and an exponential distribution with mean 1 for G_0 for our simulation studies and data analysis. Alternatively, the prior process $\pi_4(F_\epsilon)$ based on (8) is expressed as the following hierarchical structure:

$$\begin{aligned} \varepsilon_i | \sigma_i &\sim N(0, \sigma_i^2) \\ \sigma_i &\sim G \\ G &\sim DP(G_0, \alpha) \end{aligned} \quad (9)$$

where $DP(G_0, \alpha)$ is approximated via the stick-breaking process (Sethuraman, 1994) as follows:

$$\begin{aligned} G &\simeq \sum_{k=1}^K V_k \prod_{j=1}^{k-1} (1 - V_j) \delta_{\theta_k} = \sum_{k=1}^K w_k \delta_{\theta_k}, \\ V_k &\stackrel{\text{iid}}{\sim} \text{Beta}(1, \alpha), \\ \theta_k &\stackrel{\text{iid}}{\sim} G_0, \end{aligned} \quad (10)$$

where K is chosen large enough to obtain a reasonable approximation of the $DP(G_0, \alpha)$ in (9). Box and Cox (1964) recommend restricting the parameter λ of the transformation function of (3) within the interval $(0, 3]$. Accordingly, we set $\pi_3(\lambda)$ as $Unif(0,3)$ for our data analysis. We recommend that the choices for $\pi_1(\beta)$ and $\pi_2(\gamma)$ should be specific to the study. We can also use historical data, expert opinions, and other underlying biological/clinical information to specify these prior densities (Chen *et al.*, 1999). Researchers often choose normal densities as an useful class for such priors, and we have done so in Sections 4 and 5.

4 Simulation study

For our simulation studies, we use the log-median survival time of the noncured patient as $M(x) = \beta_0 + \beta_1 x = 0.1 + 0.5x$, the probability of being cured as $\eta(x) = 1/(1 + \exp(-\gamma_0 - \gamma_1 x))$ with $\gamma_0 = -1.386$, $\gamma_1 = 0.539$. For each simulated dataset, the univariate covariate x can take five possible values 0, 1, 2, 3, 4 with an equal number of patients for each value. Using the TBSCRM of (1) with monotone transformation parameter $\lambda = 2.1$, we simulate 500 datasets for each of three different sample sizes, $n = 100, 200$, and 500. The error ε of (2) is simulated from a mixture of $Unif(-1, +1)$ and $Laplace(0, 0.1)$ densities with equal weights. Survival times greater than or equal to 30 have been censored at 30.

For the simulated datasets, neither any expert opinion nor any historical data exists, so, we use a very noninformative (admittedly, somewhat impractical) $N(0, 1000)$ independent prior densities for $\beta_0, \beta_1, \gamma_0$, and γ_1 for analysis of each dataset. For each MCMC computation, we use 1000 MCMC samples, after burn-in of 10,000 samples, from each of four independent MCMC chains initiated by values sampled from the prior distributions. Inspection of these four independent MCMC chains shows that these chains converged. For the regression parameters and the transformation parameter λ , Table 1 shows Monte Carlo approximations of the coverage errors (1-coverage probability) of 95% credible intervals, the relative biases $((\hat{\theta} - \theta_0)/\theta_0)$, and the mean square errors (MSE) of point estimates and the average posterior variances. Results in Table 1 show that the Bayesian estimates of the regression parameters and λ under TBSCRM have very small relative biases, and more importantly, the approximate value of MSE (sampling expectation of square distance between estimate and parameter) is very close to the corresponding posterior variance, the estimate of MSE. We also find that the actual coverage probabilities of Bayesian credible intervals are very close to the targeted coverage probability (95%), the difference between the two decreases with increase in sample-size.

We now consider the performance of the Bayesian estimators of the following quantiles of the survival times for the five different covariate values: $Q(\tau = 0.40|x=0) = \exp(\beta_0) = 1.105$; $Q(\tau = 0.35|x=1) = \exp(\beta_0 + \beta_1) = 1.822$; $Q(\tau = 0.29|x=2) = \exp(\beta_0 + 2\beta_1) = 3.004$; $Q(\tau = 0.22|x=3) = \exp(\beta_0 + 3\beta_1) = 4.953$; $Q(\tau = 0.16|x=4) = \exp(\beta_0 + 4\beta_1) = 8.166$, where $H[T < Q(\tau|x)|x] = \tau$ is the quantile function given in (5). Table 2 presents coverage errors of 95% credible intervals, the approximate relative biases, the approximate mean square errors (MSE), and the average posterior variances of Bayesian estimates of these $Q(\tau|x)$ values based on the 500 simulated datasets for each simulation setting. Results of Table 2 show that the Bayesian estimates of the quantile functions have very small relative bias, and

the approximate MSE values are also very close to the corresponding average posterior variances (estimated MSE), indicating high level of accuracy of the posterior variance as a measure of uncertainty of the estimates. Using the stick-breaking process described in (10), the approximate Bayes estimates of the error distribution $F_\epsilon(u)$ can be computed as $\hat{F}_\epsilon(u) = (1/M) \sum_{j=1}^M \sum_{k=1}^K w_{(j)k} \Phi(u/\theta_{(j)k})$, where $\Phi(\cdot)$ is the CDF of standard normal, $(w_{(j)k}, \theta_{(j)k})$ for $j = 1, \dots, M$ are MCMC samples of (w_k, θ_k) described in (10). For three datasets of the sample sizes $n = 100, 200$, and 500 simulated from TBSCRM of (1), Fig. 1(a–c) shows the true error distributions $F_\epsilon(u)$ and the corresponding Bayesian estimates $\hat{F}_\epsilon(u)$ under two competing models: the TBSCRM and the model of Wu and Yin (2013). We refer to the Bayesian version of Wu and Yin (2013)'s model as BWYM (Bayesian Wu and Yin's Model) for the rest of the paper. The conditional survival time of an uncured patient in BWYM follows a Log-Linear model. As we can see from the plots in Fig. 1(a–c), the estimates $\hat{F}_\epsilon(u)$ from TBSCRM are close to the true $F_\epsilon(u)$, and these two curves get closer as the sample size of the data increases. It implies that our Bayesian method with TBSCRM provides very good estimates of the unknown nonparametric function $F_\epsilon(u)$. Whereas, the estimate $\hat{F}_\epsilon(u)$ obtained from BWYM is not as close to the true error distributions $F_\epsilon(u)$. It implies that the transformation $g_\lambda(\cdot)$ used in (2) is necessary for fitting a good error density when the true underlying model is TBSCRM. Especially as the sample size increases, the estimated $\hat{F}_\epsilon(u)$ from TBSCRM almost overlap with the true error distributions $F_\epsilon(u)$.

Table 3 presents the comparison of the Bayesian estimates of β_1 (one covariate effect) and $Q(\tau = 0.35|x)$ from TBSCRM with those estimates from BWYM based on the above 500 simulated datasets. For both β_1 and $Q(\tau = 0.35|x)$, the Bayesian estimates from TBSCRM have smaller coverage errors for 95% credible interval, smaller relative bias, and smaller mean square errors for Bayesian point estimates, compared to corresponding Bayesian estimates from BWYM. These results show that our Bayesian method of TBSCRM provides much more accurate estimates for the regression parameters and survival time quantiles than the Bayesian method of BWYM. In addition, compared to BWYM, the Bayesian posterior variances from TBSCRM are much closer to the MSE of the Bayesian estimates. It shows evidence of potential improvements in bias, mean-square error and estimated variance for using the TBSCRM instead of using BWYM. In summary, our simulation studies suggest that the semiparametric Bayesian estimates based on TBSCRM are precise and the Bayesian inference provides accurate and honest variance estimators.

5 Application to breast cancer data

5.1 Data analysis

To illustrate the use of TBSCRM for analyzing a real-life study, we analyze the breast cancer data from the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) database. The database contains data on a sample of women diagnosed with breast cancer in December 1988, and monitored until 2011 unless experiencing either death or censoring earlier. There are three covariates: the patient's age at diagnosis (x_1), number of primary cancers (x_2), and spread of cancer stage at diagnosis. The spread is categorical and we split it in two binary covariates with local spread (x_3) and distant spread (x_4), leaving the

regional spread as the baseline. We exclude few observations whose observed (survival or censoring) times are recorded as nonpositive, as well as any observation with missing covariate values. After some adjustments for eight duplicated observations, we have 857 observations.

The survival time of interest is the time until death due to metastasis of cancer since the first diagnosis of breast cancer. All patients who died of other causes are considered censored at their times of death. In the dataset, 276 observations of 857 are uncensored. The largest observed time is 276 months (about 23 years). It is commonplace in many clinical settings to treat patients as cured of cancer if no symptoms are recorded for five to six years. So, observation interval of 23 years is clearly adequate to produce a good estimate of the cured fraction, and this dataset is a strong candidate for an analysis based on a cure-rate survival model.

5.2 Prior distributions

We use the TBSCRM of (1) with a practical modification $S_*(t|x) = S_e\{g_\lambda(\log(t+c)) - g_\lambda(x^T \beta)\}$ of (2), where constant c ($= 1$ for the SEER data) is chosen only for the ease of the likelihood and MCMC computations. As a consequence, given $T < +\infty$ for an uncured patient, her conditional median of $\log(T+1)$ is $M = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, that is, $P[T > \exp(M) - 1 | T < +\infty; x] = 0.5$. The probability of being cured is $\eta(x)$ with $\text{logit}(\eta) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4$.

The prior distributions for λ and F_e are the same as discussed in Section 3.2. Figure 1(d) shows the plot of the Bayesian estimates of the error distribution after fitting this data. The prior distribution for β_0 is based on the prior opinion about the median survival time ($\exp(\beta_0) - 1$) of a uncured patient with $x_1^* = x_2^* = x_3^* = x_4^* = 0$. Accordingly, we chose a prior density for β_0 as $\beta_0 \sim N(4.11, (0.25)^2)$, which implies that the prior guess for an uncured patient's median survival time is 5 years with 95% prior probability of this unknown median to be between 3 and 8 years. Selection of prior distribution for β_1 is based on the prior opinion about the ratio of the medians of two uncured patients with only one unit difference in their 1st covariate values. The chosen prior of β_1 is $N(0, (0.25)^2)$, which implies that the ratio of the median survival times of two uncured patients with only one year age difference is a priori centered at $e^0 = 1$ with approximately 95% prior probability being in the range $(2e^{-0.25}, 2e^{0.25})$. The priors for $\beta_2, \beta_3, \beta_4$ are also chosen as $N(0, (0.25)^2)$ using similar arguments.

The prior distribution of γ_0 is $N(0.85, (0.33)^2)$, which implies a prior guess of a 70% cure-rate with a 99.7% prior probability of the cure-rate being between 60% and 80% for a patient with covariate values 0. Selections of prior distributions for $\gamma_1, \gamma_2, \gamma_3$, and γ_4 are based on prior opinions about the odds ratios (OR) being cured for two patients with different covariate values. For example, our prior distribution for γ_1 is $N(0, (0.29)^2)$, which implies that the OR of being cured has a prior guess of 1 and 95% prior probability of being between 0.6 and 1.8 for two patients with exactly one year age difference. The prior distributions for $\gamma_2, \gamma_3, \gamma_4$ are also set to be $N(0, (0.29)^2)$ using similar arguments.

5.3 Parameter estimates

For TBSCRM, Table 4 shows the posterior median estimates of the parameters associated with covariate effects. Figure 2(a–c) shows five quantile estimates for uncured patients with different ages, different spread groups, but all with two primary cancers. Figure 2(d) shows the effect of the number of primary cancers on the quantiles when an uncured patient's age of diagnosis is 30 years and the cancer is at the regional spread stage. The Bayesian estimated parameters of age effects indicate that diagnosis at older age would result in the smaller probability of being cured and a shorter median survival time for an uncured patient. A similar relationship between age at diagnosis and other survival time quantiles can be seen in Fig. 2(a–c). However, the credible intervals of the age effect parameters are very close to 0, showing the small effects of the age at diagnosis. The Bayesian estimates of the effects of the number of primary tumors show that the bigger number of primary tumors is expected to decrease the probability of a cure, however, suprisingly, it is expected to increase the survival time if the patient is not cured. For example, Fig. 2(d) shows that 90th percentile increases from about 300 month to about 800 month when the number of primary cancers increase from 1 to 4. This may indicate positive effects of aggressive treatment of patients with higher number of primary tumors. The parameter estimates also show that local spread has good effects, both in terms of longer survival times for uncured patients and bigger probability of cure. We found the distant spread have bad prognosis in terms of shorter survival times for uncured patients and lower probability of cure. As can be seen from Fig. 2(a–c), the median survival time is the longest when the spread stage is local and the shortest when the spread stage is distant. So, our results also make sense from a clinical viewpoint, because cancer that spreads less should be easier to treat successfully.

5.4 Model evaluation

For comparing multiple Bayesian survival models with cured fractions, Gu *et al.* (2011) proposed the M -criterion using difference between the observed cumulative incidence function $N_0(t) = \sum_{i=1}^n \delta_i I[t_i \leq t]$ from data D and the predictive path of cumulative number of incidences $N_p(t)$ simulated from the fitted Bayesian model and assuming the censoring distribution to be $\hat{F}_c$, the nonparametric Kaplan-Meier estimator of censoring distribution based on D . A smaller value of the M -criterion M_m suggests the model m being a more appropriate model for the observed data D because the predicted sample path $N_p(t)$ of cumulative incidence function under model m is closer to the observed path $N_0(t)$. An approximation of the M_m for our problem using Q MCMC samples of the posterior predictive values of $(N_p(t_1), \dots, N_p(t_n))$ is $(1/Q) \sum_{j=1}^Q M_j$ with

$$M_j = \sum_{k=n}^K (t_{(k)} - t_{(k-1)}) \{N_0(t_{(k)}) - N_{pj}(t_{(k)})\}^2, \quad (11)$$

where the vector $(N_{pj}(t_1), \dots, N_{pj}(t_n))$ is the j -th posterior predictive sample (obtained via MCMC) of $N_p(t)$ at distinct observation times $0 = t_{(1)} < \dots < t_{(n)} = \tau$ of D . Details about posterior sampling procedure of $N_p(t)$ using MCMC samples of posterior-predictive survival times and predictive censoring times (sampled from $\hat{F}_c$) are described in Gu *et al.* (2011).

Using the M -criterion scores M_m for different models, we compare the TBSCRM to two widely used existing cure-rate models. The first model is the First Activation Model (FAM) of Cooner *et al.* (2007) with the population survival function $S(t) = \exp[-\theta\{1 - S_*(t)\}]$, where $S_*(t) = \exp\{-\exp(\mathbf{x}^T \boldsymbol{\beta})t^\rho\}$ is the survival function for the latent activation time. The second model is the Proportional Odds Model (POM) of Gu *et al.* (2011) with the population survival function $S(t) = \{1 + \exp(\mathbf{x}^T \boldsymbol{\beta})F_*(t)\}^{-1}$, where the distribution of the latent survival times is $F_*(t) = 1 - \exp(-\zeta t^\rho)$. The M -criterion values for TBSCRM, FAM, and POM are 89,378, 9,624,649, and 63,676,278, respectively. The TBSCRM is the model most supported by the observed breast cancer data, as demonstrated by the smallest M -criterion score. POM is the least supported model by the observed data because it has the largest M -criterion value.

To understand more clearly how these three models perform at various observed failure time points, we compare $R_m(t)$, the posterior predicted sample path of number of patients at risk at t under model m , with $R_0(t)$, the observed count of the number of patients at risk at time t in data D . As a diagnostic plot for evaluating the appropriateness of model m , we plot

$$\tilde{R}_m(t) = \sum_{i=1}^n \hat{P}[R_i(t) = 1 | \mathbf{x}_i, m] = \hat{S}_c(t) \sum_{i=1}^n \hat{S}(t | \mathbf{x}_i, m) \quad (12)$$

along with $R_0(t)$ to evaluate the fit of the model m based on the observed data D , where $\hat{S}(t | \mathbf{x}_i, m)$ is posterior estimate of $S(t | \mathbf{x}_i, m)$ based on the posterior under model m .

Figure 3 shows the plots of $\tilde{R}_m(t)$ over time t for TBSCRM, FAM, and POM, along with $R_0(t)$ (observed) for the breast cancer data. The model with $\tilde{R}_m(t)$ closer to $R_0(t)$ indicates that model m provides more accurate estimates of the observed number of patients at risk, and thus implies a better fit. Figure 3 shows that in general, all three models appear to fit very well to the SEER data, and the TBSCRM has the best fit among all three models because the $\tilde{R}_m(t)$ obtained from TBSCRM is the closest to observed $R_0(t)$ for the entire time period. In particular, TBSCRM has the better fit compared to POM and FAM when time is less than 80 month. In this interval, $\tilde{R}_m(t)$ using models other than TBSCRM underestimate the observed $R_0(t)$. After 80 month, $\tilde{R}_m(t)$ under TBSCRM and POM are hard to distinguish and they both underestimate $R_0(t)$ in this interval. Overall, FAM based $\tilde{R}_m(t)$ has the maximum bias compared to $R_0(t)$ in the entire time range.

6 Conclusion

We present a new semiparametric cure-rate survival model called the transform-both-sides cure-rate model (TBSCRM), which models long-term cancer survival data in the presence of patients that may be cured from the disease of interest. The form of the TBSCRM provides straightforward interpretation of the covariate effects on the probability of cure as well as on the quantiles of survival time for uncured patients. Our simulation studies demonstrate the ability of the TBSCRM with Bayesian method to estimate the parameters more accurately compared to existing methods. We also use TBSCRM to present a Bayesian analysis of SEER breast cancer dataset. We show that TBSCRM outperforms the two other competing

models. It fits the SEER data better in terms of diagnostic plots as well as according to our M -criterion, extended to deal with cure-rate. Most importantly, we show that our TBSCRM provides an explicit expression of the quantile function for a patient's survival time (unlike the existing cure-rate survival models). We have presented the Bayesian estimates of the quantile functions and demonstrate their use to interpret the results of SEER breast cancer study.

Appendix:: Derivation of the quantile function

From (1), define $y = \log(t)$, then

$$P[T_i > t | \mathbf{x}_i] = S(t | \mathbf{x}_i) = \eta(\mathbf{x}_i) + \{1 - \eta(\mathbf{x}_i)\} S_\epsilon \{g_\lambda(y) - g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta})\}. \quad (\text{A1})$$

It is known that $P[T_i < Q_*(\tau) | T_i < +\infty] = \tau$, and $P(\epsilon < \epsilon_\tau^*) = \tau$. We define $\tau_p = P[T_i < Q(\tau_p)]$, then $\tau_p = \{1 - \eta(\mathbf{x}_i)\} \tau$. Therefore,

$$\tau_p = \{1 - \eta(\mathbf{x}_i)\} P(\epsilon < \epsilon_\tau^*). \quad (\text{A2})$$

From the definition, we also have $1 - \tau_p = S(Q(\tau_p) | \mathbf{x}_i)$. It is also known that $Q_*(\tau) = Q(\tau_p)$, therefore, $1 - \tau_p = \eta(\mathbf{x}_i) + \{1 - \eta(\mathbf{x}_i)\} S_\epsilon \{g_\lambda(\log Q_*(\tau)) - g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta})\}$. Incorporating (A2) to this result, we obtain that

$$1 - \{1 - \eta(\mathbf{x}_i)\} P(\epsilon < \epsilon_\tau^*) = \eta(\mathbf{x}_i) + \{1 - \eta(\mathbf{x}_i)\} S_\epsilon \{g_\lambda(\log Q_*(\tau)) - g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta})\}. \quad (\text{A3})$$

Notice that $1 - P(\epsilon < \epsilon_\tau^*) = S_\epsilon(\epsilon_\tau^*)$. Therefore, (A3) can be simplified as

$S_\epsilon(\epsilon_\tau^*) = S_\epsilon \{g_\lambda(\log Q_*(\tau)) - g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta})\}$, and, since the survival function $S_\epsilon(\cdot)$ of the error is monotone, we have that $\epsilon_\tau^* = g_\lambda(\log Q_*(\tau)) - g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta})$. After rearrange it, the quantile function $Q_*(\tau)$ is $Q_*(\tau) = \exp \{g_\lambda^{-1}(g_\lambda(\mathbf{x}_i^T \boldsymbol{\beta})) + \epsilon_\tau^*\}$.

References

- Berkson J and Gage RP (1952). Survival curve for cancer patients following treatment. Journal of the American Statistical Association 47, 501–515.
- Bickel PJ and Doksum KA (1981). An analysis of transformations revisited. Journal of the American Statistical Association 76, 296–311.
- Boag JW (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society, Series B 11, 15–44.
- Box GEP and Cox DR (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B 26, 211–243.
- Chen M, Ibrahim J and Sinha D (1999). A new Bayesian model for survival data with a surviving fraction. Journal of the American Statistical Association 94, 909–919.
- Cooner F, Banerjee S, Carlin BP and Sinha D (2007). Flexible cure-rate modeling under latent activation schemes. Journal of the American Statistical Association 102, 560–572. [PubMed: 21031152]
- Ferguson TS (1973). Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230.

- Fonseca RS, Valenca DM and Bolfarine H (2013). Cure rate survival models with missing covariates: a simulation study. *Journal of Statistical Computation and Simulation* 83, 97–113.
- Gelfand AE and Smith AFM (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398–409.
- Gu Y, Sinha D and Banerjee S (2011). Analysis of cure rate survival data under proportional odds model. *Lifetime Data Analysis* 1, 123–134.
- Hanson T and Johnson W (2002). Modeling regression error with a mixture of polya trees. *Journal of the American Statistical Association* 97, 1020–1033.
- Hjort NL and Walker SG (2009). Quantile pyramids for Bayesian nonparametrics. *The Annals of Statistics* 37, 105–131.
- Ibrahim JG, Chen M and Sinha D (2001). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics* 57, 383–388. [PubMed: 11414560]
- Kottas A and Gelfand AE (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association* 96, 1458–1468.
- Kottas A and Krnjajic M (2009). Bayesian nonparametric modeling in quantile regression. *Scandinavian Journal of Statistics* 36, 297–319.
- Lin J, Sinha D, Lipsitz S and Polpo A (2012). Semiparametric Bayesian survival analysis using models with log-linear median. *Biometrics* 59, 1–10.
- Peng L and Huang Y (2008). Survival analysis with quantile regression models. *Journal of American Statistics Association* 103, 637–649.
- Plummer M (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* Vienna, Austria.
- Portnoy S (2003). Censored regression quantiles. *Journal of American Statistics Association* 98, 1001–1012.
- Reich BJ, Bondell HD and Wang HJ (2010). Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics* 11, 337–352. [PubMed: 19948746]
- Reich BJ and Smith LB (2013). Bayesian quantile regression for censored data. *Biometrics* 69, 651–661. [PubMed: 23844559]
- Sethuraman J (1994). A constructive definition of dirichlet priors. *Statistica Sinica* 4, 639–650.
- Todkar ST and Kadane JB (2011). Simultaneous linear quantile regression: a semiparametric Bayesian approach. *Bayesian Analysis* 6, 1–12. [PubMed: 22247752]
- Tsodikov AD (1998). A proportional hazards model taking account of long-term survivors. *Biometrics* 54, 1508–1516. [PubMed: 9883549]
- Wu Y and Yin G (2013). Cure rate quantile regression for censored data with a survival fraction. *Journal of the American Statistical Association* 108, 1517–1531.
- Yakovlev AY, Asselain B, Bardou VJ, Fourquet A, Hoang T, Rochefediere A and Tsodikov AD (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *in Biometrie et Analyse de Donnees Spatio-Temporelles* 12, 66–82.
- Yin G and Ibrahim JG (2005). Cure rate models: a unified approach. *The Canadian Journal of Statistics* 33, 559–570.

Figure 1.

(a–c): Error distribution plots of one simulated dataset from TBSCRM for three different sample sizes: Distribution of error (ε) (thin solid line) and corresponding Bayesian estimates under the competing models: (1) TBSCRM (bold solid line); (2) BWYM (dashed line). (d): Error distribution plots of the breast cancer data: Bayesian estimates of the distribution of ε under the competing models: (1) TBSCRM (solid line); (2) BWYM (dashed line).

Figure 2.

Breast cancer data: (a–c), Bayesian estimates of the survival time quantiles for uncured patients versus age at diagnosis (x_1) for the three different cancer spread stage. All survival time quantiles are estimated under TBSCRM and for patients with 2 primary cancers; (d), Bayesian estimates of the survival time quantiles under TBSCRM for uncured patients whose cancer spread stage is regional and age at diagnosis of breast cancer is 30 years.

Figure 3.

Breast cancer data: counts of observed number of patients at risk over time (solid line) and corresponding Bayesian estimates under three competing models: (1) TBSCRM (bold dotted line); (2) POM (dashed line); (3) FAM (dotted line).

Table 1

Results of the simulation study under TBSCRM: coverage error of 95% credible interval, relative bias, approximate mean square error (MSE), and the associated Monte Carlo standard error (within parenthesis), and the average of posterior variance of Bayesian estimates based on 500 simulated datasets.

Sample size	Parameter	True value	Coverage error	Approximate relative bias	Approximate MSE	Average of posterior variance
100	β_0	0.1	0.084	-0.1198	0.0144 (0.0271)	0.0145
200			0.062	-0.0241	0.0034 (0.0071)	0.0033
500			0.048	0.0177	0.0007 (0.0010)	0.0007
100	β_1	0.5	0.086	0.0050	0.0010 (0.0017)	0.0010
200			0.062	0.0016	0.0002 (0.0005)	0.0002
500			0.054	-0.0007	0.0000 (0.0001)	0.0001
100	γ_0	-1.386	0.066	-0.0369	0.3135 (0.4586)	0.2873
200			0.052	-0.0483	0.1388 (0.2060)	0.1346
500			0.036	-0.0079	0.0048 (0.0693)	0.0517
100	γ_1	0.539	0.068	0.0285	0.0308 (0.0479)	0.0533
200			0.066	0.0456	0.0138 (0.0195)	0.0133
500			0.038	0.0101	0.0048 (0.0063)	0.0051
100	λ	2.1	0.068	0.0271	0.0795 (0.1357)	0.0727
200			0.052	0.0273	0.0324 (0.0485)	0.0343
500			0.040	0.0211	0.0129 (0.0181)	0.0129

Table 2

Results of the simulation study under TBSCRM: coverage error of 95% credible interval, relative bias, approximate mean square error (MSE), and the associated Monte Carlo standard error (within parenthesis), and the average of posterior variance of Bayesian estimates of selected quantiles based on 500 simulated datasets.

Sample size	True value	Coverage error	Approximate relative bias	Approximate MSE	Average of posterior variance
$Q(\tau = 0.40 x = 0)$					
100	1.105	0.084	-0.005	0.016 (0.027)	0.017
200		0.062	0.001	0.004 (0.010)	0.004
500		0.048	0.002	0.001 (0.001)	0.001
$Q(\tau = 0.35 x = 1)$					
100	1.822	0.068	-0.005	0.026 (0.044)	0.026
200		0.064	0.000	0.006 (0.015)	0.006
500		0.042	0.002	0.001 (0.002)	0.001
$Q(\tau = 0.29 x = 2)$					
100	3.004	0.070	-0.005	0.034 (0.059)	0.036
200		0.060	0.000	0.008 (0.018)	0.008
500		0.038	0.001	0.002 (0.002)	0.002
$Q(\tau = 0.22 x = 3)$					
100	4.953	0.038	-0.004	0.036 (0.062)	0.081
200		0.056	0.000	0.009 (0.018)	0.010
500		0.030	0.001	0.002 (0.002)	0.002
$Q(\tau = 0.16 x = 4)$					
100	8.166	0.040	-0.001	0.070 (0.109)	0.812
200		0.050	0.001	0.019 (0.032)	0.026
500		0.016	0.000	0.003 (0.006)	0.005

Table 3

Results of the simulation study to compare Bayesian TBSCRM with BWYM (Bayesian Wu and Yin's model) for estimation of $\beta_1 = 0.5$ and $Q(\tau = 0.35|x = 1) = 1.822$ using 500 simulated datasets from TBSCRM for three different sample sizes.

Model	Sample size	Coverage error	Relative bias	Approximate MSE	Average of posterior variance
$\beta_1 = 0.5$					
TBSCRM	100	0.086	0.005	0.001 (0.002)	0.001
	200	0.062	0.002	0.000 (0.001)	0.000
	500	0.054	-0.001	0.000 (0.000)	0.000
BWYM	100	0.108	-0.015	0.005 (0.029)	0.004
	200	0.118	-0.009	0.003 (0.026)	0.001
	500	0.182	0.007	0.005 (0.036)	0.000
$Q(\tau = 0.35 x = 1) = 1.822$					
TBSCRM	100	0.068	-0.005	0.026 (0.044)	0.026
	200	0.064	0.000	0.006(0.015)	0.006
	500	0.042	0.002	0.001 (0.002)	0.001
BWYM	100	0.166	0.094	0.684 (9.397)	1.069
	200	0.192	0.044	0.028(0.117)	0.032
	500	0.266	0.038	0.036 (0.299)	0.009

Notes

* Coverage error: coverage error of 95% credible interval estimates.

* Relative bias: average relative bias of the Bayesian posterior median estimates over 500 datasets.

* Approximate MSE: average mean square error of the Bayesian posterior median estimates over 500 datasets. Within parathesis are the Monte Carlo standard errors of the MSEs.

* Average of posterior variance: average variance of Bayesian posterior samples over 500 datasets.

Table 4

Breast cancer data: pointwise and 95% interval estimates of the model parameters under TBSCRM.

Covariate	Survival model parameters		Cure rate parameters	
	Estimates	95% Interval	Estimates	95% Interval
Intercept	3.919	(3.502, 4.355)	0.959	(0.408, 1.508)
Age	−0.000	(−0.008, 0.009)	−0.007	(−0.020, 0.005)
Primaries	0.324	(0.086, 0.542)	−0.598	(−1.045, −0.213)
Local spread	0.739	(0.413, 1.074)	1.069	(0.609, 1.521)
Distant spread	−0.942	(−1.282, −0.598)	−1.095	(−1.742, −0.447)