Published in final edited form as:
Lifetime Data Anal. 2011 January ; 17(1): 123–134. doi:10.1007/s10985-010-9171-z.

Analysis of Cure Rate Survival Data Under Proportional Odds Model

Yu Gu[1,*], **Debajyoti Sinha**[1,**], and **Sudipto Banerjee**[2,***]

[1]Department of Statistics, Florida State University, Tallahassee, Florida 32310–5608, U.S.A.

[2]MMC 303, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455–0392, U.S.A.

SUMMARY

Due to significant progress in cancer treatments and management in survival studies involving time to relapse (or death), we often need survival models with *cured fraction* to account for the subjects enjoying prolonged survival. Our article presents a new *proportional odds* survival models with a cured fraction using a special hierarchical structure of the latent factors activating cure. This new model has same important differences with classical proportional odds survival models and existing cure-rate survival models. We demonstrate the implementation of Bayesian data analysis using our model with data from the SEER (Surveillance Epidemiology and End Results) database of the National Cancer Institute. Particularly aimed at survival data with cured fraction, we present a novel Bayes method for model comparisons and assessments, and demonstrate our new tool's superior performance and advantages over competing tools.

Keywords

Bayesian hierarchical models; Cure rate models; Cure fractions; Latent activation schemes; Markov Chain Monte Carlo algorithms; Proportional Odds Models; Survival analysis

1. Introduction

With rapid developments in medical and health sciences, we now encounter more survival studies where some patients are expected to be *cured*. Survival models that account for cure are important for understanding prognosis in potentially terminal diseases. Traditional parametric survival models such as Weibull or Gamma (see, e.g., Cox and Oakes, 1984) do not account for the probability of cure. Although subtle, one needs to distinguish between the concepts of censoring and cure: censoring refers to a subject who does not fail within the monitoring time window of a particular subject, while cure refers to one who will not fail within any reasonable monitoring time window. Indeed the latter is an abstraction as we never "observe" a cure (due to a finite monitoring time). Still estimating the probability of such an outcome, especially in various cancer-relapse settings, can help expose unknown health issues concerning that population.

Recently much attention has been devoted to formulating parametric survival models incorporating a *cured fraction* – a non-zero tail probability of the survival function. These have focused upon cancer-relapse trials including breast cancer, non-Hodgkins lymphoma,

*yugu@stat.fsu.edu. **sinhad@stat.fsu.edu. ***sudiptob@biostat.umn.edu.

leukemia, prostate cancer, melanoma, and head and neck cancer, where due to recent advances in therapy and treatment, a significant proportion of patients are expected to be "cured", that is to remain disease-free even after really long follow-ups. *Cure rate models* incorporating a cured fraction, defined as a non-zero tail-probability of the survival function, adjust for this feature of the data and date back to the mixture model by Berkson and Gage (1954) (BG model, in short) and has been extensively discussed by several authors, including Farewell (1982, 1986), Gray and Tsiatis (1989), Maller and Zhou (1996), Ewell and Ibrahim (1997), Stangl and Greenhouse (1998), and Sy and Taylor (2000). In this model, the survivor function for the entire population is given by

$$P[T>t]=S_p(t)=\pi+(1-\pi)\,S(t), \tag{1.1}$$

where $\pi = S_p(+\infty)$ is the "cured fraction", and $S(t)$ with $S(+\infty) = 0$ is the *proper* survivor function for the non-cured group. In the presence of the $p \times 1$ vector of covariates $x_i'=(x_{i1},\ldots,x_{ip})$ for the i^{th} subject, assuming an accelerated failure time model $S(t|x_i) = S_0(t\theta(x_i))$ for non-cured subject and the cured fraction π to be free of x_i, we get obtain an accelerated failure time model

$$S_p(t|x_i)=\pi+(1-\pi)S_0(t\theta(x_i))=S_{p0}(t\theta(x_i)) \tag{1.2}$$

for the population survival function.

Another class of models, formulated by Yakovlev et al. (1993), Yakovlev (1994), Yakovlev and Tsodikov (1996) and Chen, Ibrahim and Sinha (1999) (YCIS model, in short) in cancer relapse settings, assume that a latent biological process of propagation of latent clonogenic tumor cells (latent factors) is generating the observed failure (relapse). Cooner et al. (2007) generalized this framework to a flexible class of cure models under latent activation schemes. Consider a typical cancer setting where for each individual in the population under study, we posit a certain unknown number, N, of latent factors. Let Z_i be the time (promotion time) for the i^{th} latent factor. Given $N > 0$, $Y_1, \ldots, Y_N$ are assumed to be independent and identically distributed with a common distribution function $F(y) = 1-S(y)$ that does not depend upon N. The time to final event of interest can be defined by $T = \min\{Y_i : 1 \le i \le N\}$, when $N > 1$. This model can be used in cancer relapse or other disease models whenever we can envisage one or several *latent factors* or *events* corresponding to each patient. For an individual to be at *risk* of failure, he/she must be exposed to at least one of these latent factors. If $N = 0$, then the individual is not at risk of final event and is considered *cured*. Failure is observed when one (or some) of these latent factors get *activated/prommoted*.

Notice that N must be modeled using a stochastic mechanism. The number of possible latent events N can have any finite-mean integer-valued distribution (e.g., Binary, Geometric, etc.) with the moment generating function $m(t) = E[\exp(tN)]$ and a *cure fraction* defined as $P(N = 0) = m(-\infty)$. In this setting, the marginal distribution of T is given in terms of $m(t)$ as (Cooner et al., 2007):

$$S_p(t)=E_N[P(T \ge t|N)]=m[\log S(t)]. \tag{1.3}$$

For example, in the traditional BG model, N is binary $N \sim \text{Ber}(\theta)$ $(0 \le \theta \le 1)$ with $m(t) = 1 - \theta(1 - e^t)$ to give

$$S_p(t)=(1-\theta)+\theta S(t), \tag{1.4}$$

with cure fraction $1-\theta \leq 1$. The YCIS model assumes that N has a Poisson distribution with $m(t) = \exp[-\theta(1-e^t)]$ for $\theta > 0$ and the corresponding marginal cure rate model is

$$S_p(t)=e^{-\theta(1-S(t))}, \tag{1.5}$$

with cure fraction $\exp(-\theta)$. The biological arguments for using this assumption for a cancer relapse study are put forward by Hanin et al. (2001) and Hanin (2001) among others. However, Tucker and Taylor (1996), among others, find the Poisson assumption at best debatable irrespective of any situation involving cure in cancer. The class of models in (1.3) is far more general than these two competing models in existing literature of cure-rate survival data.

The hazard function of the YCIS model in (1.5) is given by

$$h_p(y)=\theta f(y), \tag{1.6}$$

when the covariate vector $\boldsymbol{x}_i$ for the i^{th} subject is incorporated through the cure rate parameter θ_i as $\theta_i \equiv \theta(\boldsymbol{x}_i'\beta) = \exp(\boldsymbol{x}_i'\beta)$, where $\beta = (\beta_1, \ldots, \beta p)'$ denote the corresponding vector of regression coefficients, and $F(t)$ is assumed to be free of $\boldsymbol{x}_i$ to get a proportional hazards structure for the population hazard in (1.6).

Most of the existing cure models in the literature are modifications of either the BG (see, e.g., Sy and Taylor, 2000, Li and Taylor, 2002, Banerjee and Carlin, 2004) or the YCIS models (see, e.g., Tsodikov et al., 2003). Our first goal here is to develop another class of cure-rate models where the survival function $S_p(t|\boldsymbol{x})$ will have a proportional odds structure, an already popular regression model in ordinary survival model (see, e.g., Bennett, 1983 and Collett, 1994). Our second goal is to present the associated Bayesian method including a novel Bayesian model diagnostic tool for cure-rate survival data. The cross-validated method like L-criterion (e.g. Ibrahim, Chen and Sinha, 2001) popular for uncensored data from linear and generalized linear model depends on difference between observed and predicted, and their extension is not obvious when observed data is subject to right censoring. The Bayesian cross-validated predictive density based method of CPO (e.g., Ibrahim et al., 2001b) is problematic particularly for cure-rate model because the CPO for a subject can be either density-value (for uncensored non-cured) or probability (for either censored or cured).

In Section 2, as an alternative to the YCIS and BG models, we develop a new general class of cure models having a proportional odds structure. The key feature of the proportional odds survival model with cured fraction is that the ratio of hazards for two covariate values does not remain constant over time (unlike the Cox model structure of the YCIS). However, unlike the hazard ratio for a proportional odds survival model with no cured fraction, the hazard ratio for the cure-rate proportional odds model does not go to one at infinity. We investigate the key characterizations and properties of this model. Specifically, we show how this cure-rate proportional odds model can be characterized by the latent factors model of Cooner et al. (2007) with a geometric distribution for the number of latent factors. We also discuss the posterior and Bayes estimation of our model. Section 3 introduces our Bayesian method of model selection particularly aimed at right-censored survival data.

Section 4 illustrates the performance of our models with the melanoma data. Section 5 concludes the paper with a summary and indicates future areas of related research.

2. Models, Properties and Posteriors

The cure models envisaged by Yakovlev (1994) and others deal with failure (relapse) times at two different levels: an *observed* failure time, say T, corresponding to the time when the individual *fails*, and the *latent* event times, Y_k, $k = 1, \ldots, N$, the activation times for the N latent factors that generate the observed failure at time T. Note that if $N = 0$ then the individual is not exposed to any of the latent factors and is considered immune from failure. Conditional upon N, the Y_k's are assumed to be independently and identically distributed with a *latent survival function* $P(Y > t) = S(t) = 1-F(t)$. When N is distributed as $Geo(\theta)$ with p.d.f. $P[N = k] = \theta^k/(1 + \theta)^{k+1}$, then we get the population survival function as

$$S_p(t)=[1+\theta F(t)]^{-1}. \tag{2.7}$$

This survival function in (2.7) has proportional odds structure when covariate $\boldsymbol{x}$ is modeled via $\theta(\boldsymbol{x})$ and the latent survival $S(t) = 1 - F(t)$ is free of x, because

$$\{1 - S_p(t|\boldsymbol{x})\}/S_p(t|\boldsymbol{x})=\theta(\boldsymbol{x})F(t). \tag{2.8}$$

The corresponding hazard is

$$h_p(t|\boldsymbol{x})= -\frac{d}{dt}\log\{S_p(t|\boldsymbol{x})\}=\frac{\theta(\boldsymbol{x})f(t)}{1+\theta(\boldsymbol{x})F(t)}, \tag{2.9}$$

where the density $f(t)$ of $F(t)$ is assumed to be continuous except at finite time points. For proportional odds model with no cured fraction, the ratio of hazards $h(t|\boldsymbol{x}_1)/h(t|\boldsymbol{x}_2)$ goes to one as $t \to \infty$ and it goes to $\theta(\boldsymbol{x}_1)/\theta(\boldsymbol{x}_2)$ as $t \to 0$. However, for the proportional odds-model with cured fraction in (2.7), the ratio $h_p(t|\boldsymbol{x}_1)\, h_p(t|\boldsymbol{x}_2)$ goes to $[1+\theta(\boldsymbol{x}_2)]\theta(\boldsymbol{x}_1)/\{[1+\theta(\boldsymbol{x}_1)]\theta(\boldsymbol{x}_2)\}$ as $t \to \infty$, because $F(+\infty) = 1$.

We specify the latent survival function $S(t)$ using a two-parameter Weibull distribution $Weib(\rho,\eta)$ with survival function $S(t) = \exp(-\eta t^{\rho})$. This implicitly assumes that hazard $h(t) = \eta\rho t^{\rho-1}$ is either increasing (for $\rho \geq 1$) or decreasing (for $\rho \leq 1$). However, the corresponding $h_p(t|\boldsymbol{x})$ may not have the same monotonic trend. When $h(t) = \eta$ (constant), the corresponding $h_p(t|\boldsymbol{x}) = \{\theta(\boldsymbol{x})\eta\} / \{(\theta(\boldsymbol{x}) + 1)e^{\eta t} - \theta\}$ is strictly decreasing.

For the i^{th} individual, our observed data $D_i = \{y_i, \delta_i, \boldsymbol{x}_i\}$ consists of covariate vector $\boldsymbol{x}_i$, $y_i = \min(T_i, C_i)$ as the observed failure time, $\delta_i = I[T_i \leq C_i]$ as the failure indicator, where C_i is the non-informative random censoring time. We denote the model parameters (and hyper-parameters) into Ω, which actually depends on the specific model. The contribution of subject i to the data likelihood (in a right-censored setting) is

$$L(\Omega|D_i)=S_p(y_i|\Omega;\boldsymbol{x}_i) \times \{h_p(y_i|\Omega;\boldsymbol{x}_i)\}^{\delta_i}$$

where for the proportional odds model with cure rate, $S_p(t \mid \Omega; \boldsymbol{x}_i)$ and $h_p(t \mid \Omega; \boldsymbol{x}_i)$ are given in (2.7) and (2.9) respectively. For other models, such as the BG and the YCIS model, the S_p and h_p will be corresponding to the chosen model.

The posterior distribution of Ω is

$$p(\Omega|D) \propto \left[\prod_{i=1}^{n} L(\Omega|D_i)\right] \times \pi(\Omega), \tag{2.10}$$

where $D=\{D_i\}_{i=1}^{n}$ denotes the observed data and π(Ω) is the joint prior of Ω. For the model in (2.9), it is assumed to be $\pi(\Omega) = \rho_1(\rho,\eta)\pi_2(\boldsymbol{\beta}|\rho,\eta)$. A more precise notation would acknowledge L and Ω to depend on the model m, but we supress the dependence of L and Ω on m in the notation for ease of presentation. In general the marginalization of $p(\Omega \mid D)$ is analytically intractable and is performed using Markov Chain Monte Carlo tool, which iteratively samples from the joint posterior using possibly Metropolis updates for the full conditionals. In general, we adopt normal proposals for **β**, log-normal for η and Gamma for ρ (as the case may be).

3. Bayesian Model Comparisons

With the broad range of models we can entertain for cure-rate survival data, model selection becomes an important question. Although several models may provide adequate fit to the data, each model for cure-rate survival data represents a hypothesis (or a set of hypotheses) for the mechanism of relapse, and it is beneficial to have a framework for choosing between these models. One decision-theoretic criteria proceeds from a posterior predictive loss paradigm (Gelfand and Ghosh, 1998), stating that preferred models will perform well under a decision-theoretic *balanced loss function* that eventually yields a model selection metric called the L-measure (Ibrahim, Chen and Sinha, 2001), given as

$$L_m = E\left[\sum_{i=1}^{n} (\log(t_i) - \log(t_{ip}))^2 | D\right],$$

can be used to compute the MC approximation of L_m as

$$L_m \approx \frac{1}{G}\sum_{j=1}^{G}\sum_{i=1}^{n}\left(\log(t_i) - \log\left(t_{ip(j)}\right)\right)^2,$$

where expectation is taken with respect to the posterior predictive distribution of T_{ip} for patient i given by

$$p(t_{ip}|D) = \int f(t_{ip}|\Omega)\, p(\Omega \mid D) d\Omega, \tag{3.11}$$

and $f(t_{ip} \mid \Omega)$ is the sampling density for patient i conditional upon Ω being known.

Computing (3.11) proceeds using composition sampling: given samples $\{\Omega_{(j)}\}_{j=1}^{G}$ from the posterior distribution (2.10), we sample $t_{ip(j)}$ from $f(t_{ip} \mid \Omega = \Omega_{(j)})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, G$. The samples $\{t_{ip(j)}\}_{j=1}^{G}$ from the posterior predictive distribution of the i-th subject, viz. $p(t_{ip} \mid D)$.

If there is no cure-rate and no right-censoring, the sampling of finite predictive survival times $t_{ip(j)}$ given Ω_j (MCMC sample from posterior) can be easily done. For the cure-rate survival model subject to random censoring C_i, the response consists of the pair $Y_i =$

$\min(T_i, C_i)$ and censoring indicator $\delta_i = I[T_i \le C_i]$. Therefore, we need a new definition of T_{ip} from this situation.

We propose a new measure of model diagnostic for censored data. For censored data, since the censoring indicator is also part of the observed data, the closeness between the observed $Y_i = \min(T_i, C_i)$ and the predicted Y_{ip} is not appropriate. For example, when $\delta_{obs} = 1$ and observed $y_{obs} = t_{obs} = 10$, then prediction of censoring at $y_p = 11$ is not as good as failure/death at $y_p = 11$, although $(y_{obs} - y_p)^2$ are the same for both cases. In presence of censoring and cured fraction (i.e. $E(T)$ is not finite), it may not be appropriate to perform selection based on either $E(T_i - T_{ip})^2$ or $E(Y_i - Y_{ip})^2$. To avoid any of these problems, we use the counting process of number of deaths over time to compare with the number of observed deaths over time to define our measure of model adequacy. We define the counting process $N_+(t) = \sum_{i=1}^{n} N_i(t)$, where $N_i(t) = I[y_i \le t, \delta_i = 1]$. The $N_+(t)$ denotes the number of observed failures before time t. Let $N_o(t) = \sum_{i=1}^{n} \delta_i I[y_i \le t]$ be the observed value of the counting process $N_+(t)$ from observed data D. Let $N_p(t)$ denote the posterior predicted sample path of the counting process $N_+(t)$. Our new measure, which we call the *M-measure* for model m, is defined as

$$M_m = E\left[\int_0^{\tau} |N_o(t) - N_p(t)| dt \,|\, D, F_c = \widehat{F}_{CKM}\right], \tag{3.12}$$

where $\tau = \max_i\{y_i\}$ and the distribution of censoring variable C_i is assumed to be known as $\hat{F}_{CKM}$, the Kaplan-Meier estimator of the cumulative distribution function of C from D. Different forms of norm can be used in the formula, such as the absolute value and the square.

The *M-measure* can be computed with a two-step procedure. First, we sample (using MCMC) Ω_j for $j = 1, \ldots, N$ from the posterior density $p(\Omega \mid D)$ for model m. Then, for $j = 1, \ldots, N$, the $\{y_{ipj}, \delta_{ipj}\}$ are simulated from the predictive distribution of (Y_{ip}, δ_{ip}) assuming the distribution of C to be $\hat{F}_{CKM}$. Nest, for each C_{ipj} sampled from the Kaplan-Meier cumulative density function $\hat{F}_{CKM}$, we sample $\delta_{ipj} \sim \text{Ber}(F_p(C_{ipj}))$ where $F_p(C_{ipj}) = \theta_i F(C_{ipj})/(1 + \theta_i F(C_{ipj}))$ is the population c.d.f. and set $y_{ipj} = C_{ipj}$ if $\delta_{ipj} = 0$. When $\delta_{ipj} = 1$, we sample $U_{ij} \sim U(0, 1)$ and set $y_{ipj} = F_{cure}^{-1}(U_{ij})$, where

$$F_{cure}(y) = \frac{F(y)(1+\theta_{ij}F(C_{ipj}))}{(1+\theta_{ij}F(y))F(C_{ipj})}$$

for $0 < y < C_{ipj}$ and $\theta_{ij} = \exp\left(x_i' \beta_j\right)$. Once the posterior predictive samples are obtained, for each j we compute

$$M_j = \int_0^{\tau} |N_o(t) - N_{pj}(t)| dt = \sum_{j'=1}^{K} |N_o(a_{j'}) - N_{pj}(a_{j'})| \Delta_{j'},$$

where $0 = a_0 < a_1 < a_2 < \ldots < a_K < a_{K+1} = \tau$ are distinct points where $N_o(t)$ and $N_{pj}(t)$ have jumps, and $\Delta_{j'} = a_{j'+1} - a_{j'}$. Finally, M_m is obtained as an average over the M_j's for $j = 1, \ldots, N$, where N is chosen large enough to achieve a desired level of Monte Carlo error.

Smaller values of M_m indicate better fit to the observed data as well as more precise predictive fit for the model. The advantage of this criterion, compared to *deviance information criterion* (DIC), is that the measure is based on the notion of predictive loss paradigm (e.g. Gelfand and Ghosh, 1998, Ibrahim et al., 2000), and only a very weak assumption about censoring is made for the computation of M-measure. The Kaplan-Meier estimator of the cumulative distribution function of C is not necessary. Other assumptions of the censoring time can also be used, such as the exponential distribution with the rate that has a Gamma distribution. So a wide range of models can be compared according to the M-measure.

4. Illustration With Data Analysis

4.1 Analysis of breast-cancer data

To illustrate our new methods, we analyze this breast cancer data set provided by the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) data about a national cohort of women who have been monitored for assessing breast cancer prognosis. For each of 305 patients in this database, we have racial information, age at diagnosis, the number of primary cancers that has been diagnosed, and the stage of the disease (local, regional or distant, with local as baseline). The event of interest here is the time to death since the diagnosis of first breast cancer. Only 102 patients are observed to experience the event. The other patients are considered censored, including some who might have died either from other types of cancer or other diseases. The maximum observed failure time is 84 months. With such a long observation period, the cure rate model assumption is deemed reasonable for this data set.

The prior is selected based upon the prediction for patients with median age at diagnosis after standardization 0.1133, 1 primary cancer that has been detected, and the stage of the disease as local. To show the priors are noninformative, we simulated 5 copies of observed survival time $\{Y_{i,pred} = min\{T_i, C_i\}, i = 1, \ldots, 100\}$ from the prior predictive density under model 1 where the censoring time $C_i's$ are generated from the Kaplan-Meier estimator of the original data set. Figure 1 illustrates the boxplots of the 5 copies of simulated survival time we generated based on the prior and the histogram of one copy. The medians of the 5 copies are all around the median of the observed survival time in the original data set and for each copy, it covers all the range of observed survival time in the original data set. Figure 2 is the histogram of the cured fraction we generated based on the prior. The prior median of the cured fraction is around 0.2.

Table 1 below provides the results of our Bayesian model comparison based on the M-measure and the DIC. Model 1 is the proportional odds model with cure-rate given in (2.6). Model 2 is the YCIS model, with regression on the mean of Poisson distribution. This gives us a proportional hazard structure for $h_p(t|\boldsymbol{x}) = \theta(\boldsymbol{x})f(t)$, where $\theta(\boldsymbol{x}) = e^{\boldsymbol{x}'\boldsymbol{\beta}}$ and $f(t)$ is the density function of Weibull distribution. Whereas model 3, the survival time has a Logistic distribution with regression on the location parameter μ, with $S_p(t|\boldsymbol{x}) = 1/(1 + e^{\tau(t-\mu)})$. Table 1 represents the results of different model selection methods for these three models. The proportional odds model with cure-rate (Model 1) has the lowest M-measure value which is obtained by the sampling approach of Section 3. We get consistent result when we replace the absolute value with the square. The DIC is also the lowest for Model 1 but $-2\log(CPO)$ for Model 2 is slightly smaller than that for the proportional odds model with cure-rate (less than 1%). The appropriateness of using DIC and $-2\log(CPO)$ is still questionable for censored data, but the lowest M-measure value shows the best predictive fit of Model 1 among these three models.

Figure 3 above gives a graphical comparison among the three models. We generated 500 realizations of cumulative counting processes $N_p(t)$ from the posterior predictive samples according to the method of Section 3. At each jump point a of the observed cumulative counting process $N_o(t)$, the 2.5% and 97.5% quantiles of the sampled values of all the predicted counting processes $N_o(a)$ are plotted. The observed cumulative counting process is right in the middle of the credible region in the first plot of Figure 3 while in the second or third plot, it either deviates from the center of the interval or stays out of the region when the survival time is small. Although M-measure shows that Model 1 fits the data set better, the two plots on the right in Figure 3 indicate that either Model 2 or 3 is not a very bad choice for this data set as far prediction of death due to relapse of breast cancer is concerned.

Table 2 contains the estimates from the best model we have, proportional odds model with cure rate. The four covariates are age at diagnosis, the number of primary cancers that has been detected, and the two variables for the stage of the disease with local as baseline. 0 is contained in the 95% credible intervals of the estimated coefficients for the three covariates, age and two variables for the stage of the disease, which means they are all significant. The odds ratio is multiplied by 1.6352 for every one unit of increase after standardization with a 95% credible interval (1.2937, 2.0927). The odds ratio increases by 73% if the stage of disease for a patient moves from local to reginal, with a 95% credible interval (1.0414, 2.7138). If the stage moves from local to distant, the odds ratio will be multiplied by 16.2125 with a 95% credible interval (7.1492, 39.5154). The probability of cure of a hypothetical patient with the covariates at their sample medians, (0.1133, 1, 0, 0), is estimated as 0.3035 with a 95% C.I. (0.038, 0.6717) in the last row of Table 2. This 95% credible interval shows a strong data evidence of a significant portion of cure-rate for this study.

5. Summary

We have proposed a new general class of cure rate models with a proportional odds structure. Our models keep all the advantages of regular proportional odds model for survival analysis. We can also derive our proportional odds model with cure rate from the latent factors model of Cooner et al. (2007), which makes our models share some of the properties from there. To compare the performance of our models, we developed a Bayesian model selection method that takes the censoring into consideration. Posterior predictive loss is computed as the criteria by Markov Chain Monte Carlo samples. Applying our model selection method to the breast cancer data set reveals that our proportional odds model with cure rate fits adequately compared to the other two competing models.

To extend our work, a power prior based on stage-0 data can be added to our current model. Also, it is common for patients to be at risk of death from multiple competing causes, like breast cancer and other cancers, strokes, etc. We can accommodate this feature, i.e., that one of the causes may have a cured fraction for the cause-specific survival function. For our Bayesian model selection method, theoretical explorations and simulation experiments will be needed to generalize the use of the *M-measure* to a wider range of survival data sets.

References

Banerjee S, Wall M, Carlin BP. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. Biostatistics 2003;4:123–142. [PubMed: 12925334]

Berkson J, Gage RP. Survival curve for cancer patients following treatment. Journal of the American Statistical Association 1952;47:501–515.

Chen M-H, Ibrahim JG, Sinha D. A new Bayesian model for survival data with a surviving fraction. Journal of the American Statistical Association 1999;94:909–919.

Chen M-H, Ibrahim JG, Sinha D. Bayesian inference for multivariate survival data with a cure fraction. Journal of Multivariate Analysis 2002;80:101–126.

Collett, D. Modelling Survival Data in Medical Research. London: Chapman and Hall; 1994.

Cooner F, Banerjee S, Carlin BP, Sinha D. Flexible cure rate modeling under latent activation schemes. Journal of the American Statistical Association 2007;102:560–572. [PubMed: 21031152]

Cox, DR.; Oakes, D. Analysis of Survival Data. London: Chapman and Hall; 1984.

Ewell M, Ibrahim JG. The large sample distribution of the weighted log-rank statistic under general local alternatives. Lifetime Data Analysis 1997;3:5–12. [PubMed: 9384622]

Farewell VT. The use of mixture models for the analysis of survival data with long term survivors. Biometrics 1982;38:1041–1046. [PubMed: 7168793]

Farewell VT. Mixture models in survival analysis: Are they worth the risk? Canadian Journal of Statistics 1986;14:257–262.

Gelfand AE, Ghosh SK. Model choice: a minimum posterior predictive loss approach. Biometrika 1998;85:1–11.

Gray RJ, Tsiatis AA. A linear rank test for use when the main interest is in differences in cure rate. Biometrics 1989;45:899–904. [PubMed: 2790128]

Hanin LG, Zaider M, Yakovlev AY. Distribution of the number of clonogens surviving fractionated radiotherapy: A long-standing problem revisited. International Journal of Radiation Biology 2001;77:205–213. [PubMed: 11236927]

Ibrahim JG, Chen M-H, Sinha D. Bayesian semiparametric models for survival data with a cure fraction. Biometrics 2001a;57:383–388. [PubMed: 11414560]

Ibrahim, JG.; Chen, M-H.; Sinha, D. Bayesian Survival Analysis. New York: Springer-Verlag; 2001b.

Kass RE, Raftery AE. Bayes factors. J. Amer. Statist. Assoc 1995;90:773–795.

Le, CT. Applied Survival Analysis. New York: Wiley; 1997.

Li Y, Ryan L. Modeling spatial survival data using semiparametric frailty models. Biometrics 2002;58:287–297. [PubMed: 12071401]

Li C, Taylor JMG. A semi-parametric accelerted failure time cure model. Statistics in Medicine 2002;21:3235–3247. [PubMed: 12375301]

Maller, RA.; Zhou, S. Survival analysis with long-term survivors. New York: Wiley; 1996.

Stangl DK, Greenhouse JB. Assessing Placebo response using Bayesian hierarchical survival models. Lifetime Data Analysis 1998;4:5–28. [PubMed: 9567053]

Sy JP, Taylor JMG. Estimation in a proportional hazards cure model. Biometrics 2000;56:227–336. [PubMed: 10783800]

Tsodikov AD, Ibrahim JG, Yakovlev AY. Estimating cure rates form survival data. Journal of the American Statistical Association 2003;98:1063–1078. [PubMed: 21151838]

Tucker SL, Taylor JMG. Improved models of tumour cure. Int. J. Radiat. Biol 1996;70:539–553. [PubMed: 8947535]

Yakovlev AY. Letter to the editor. Statistics in Medicine 1994;13:983–986. [PubMed: 8080583]

Yakovlev, AY.; Asselain, B.; Bardou, VJ.; Fourquet, A.; Hoang, T.; Rochefediere, A.; Tsodikov, AD. A simple stochastic model of tumor recurrence and its applications to data on pre-menopausal breast cancer. In: Asselain, B.; Boniface, M.; Duby, C.; Lopez, C.; Masson, JP.; Tranchefort, J., editors. Biometrie et Analyse de Dormees Spatio-Temporelles. Vol. 12. France: ENSA Reneed; 1993. p. 66-82.

Yakovlev, AY.; Tsodikov, AD. Stochastic models of tumor latency and their biostatistical applications. New Jersey: World Scientific; 1996.

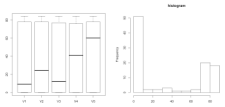

Figure 1.
Boxplots of 5 copies of simulated data sets (n=100, covariate set at median values) from prior model, and a histogram of one simulated data set.

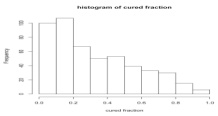

Figure 2.
Histogram of the marginal prior density of the cured fraction (for n=100, covariate evaluated at median).

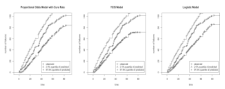

Figure 3.
A plot of the 2.5% and 97.5% quantiles of the posterior predicted cumulative counting process and the observed cumulative counting process for the breast cancer data.

Table 1

Model comparison results for breast cancer data.

Model	M-measure	DIC	−2log(CPO)
Proportional Odds Model with cure-rate	569.4	1130.886	1227.642
YCIS Model	610.8	1207.063	1193.512
Logistic Model	600.5	1685.055	1685.212

Table 2

Posterior estimates of the odds ratios and the cured fraction (evaluated at median value of covariate) using the proportional odds model with cured fraction for breast cancer data.

Regression Parameters	Median	Standard deviation	95% C.I.
Age	1.6352	3.003	(1.2937, 2.0927)
Primaries	1.132	1.2457	(0.8247, 2.0763)
Stage 1	1.7369	1.263	(1.0414, 2.7138)
Stage 2	16.2125	1.27	(7.1492, 39.5154)
Cured fraction	0.3035	0.194	(0.038, 0.6717)