Current Research in Biostatistics

Original Research Paper

An Extended Semi-Parametric Accelerated Failure Time Cure Model for Partial Cure Information Known

[1]Yu Wu, [2,3]Yong Lin, [2,3]Shou-En Lu, [4]Chin-Shang Li and [2,3]Weichung Joe Shih

[1]CR Medicon, Inc., 35 Wills Way, Piscataway, NJ 08854, USA
[2]Department of Biostatistics, School of Public Health, Rutgers,
The State University of New Jersey, 683 Hoes Lane West, Piscataway, NJ 08854, USA
[3]Division of Biometrics, Rutgers Cancer Institute of New Jersey, Rutgers,
The State University of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08901, USA
[4]School of Nursing, The State University of New York,
University at Buffalo, 3435 Main Street, Buffalo, NY14214, USA

Article history
Received: 20-08-2018
Revised: 24-08-2018
Accepted: 08-12-2018

Corresponding Author:
Yong Lin
Department of Biostatistics, School of Public Health, Rutgers, The State University of New Jersey, 683 Hoes Lane West, Piscataway, NJ 08854, USA
Email: linyo@sph.rutgers.edu

Abstract: Cure model is a useful model for analyzing failure time data when there is evidence of long-term survivors. In traditional cure models, it is assumed that the cured or uncured status in the censored set cannot be distinguished. However, in many occasions, data of some diagnostic procedures, with some sensitivity and specificity, may have provided partial information about the cured or uncured status in the censored set. Failure to use such data would be wasteful and result in efficiency loss. Wu *et al*. in 2014 proposed an extended cure model. It incorporates such additional diagnostic information into traditional Proportional Hazards (PH) cure model analysis. In this work, we extended a semi-parametric Accelerated-Failure-Time (AFT) cure model to incorporate the additional diagnostic information because AFT model may be more appropriate than PH models in some applications and it provides intuitive and easy-to-understand interpretation through postulating direct relationship between failure-times and covariates. Through simulations, we showed that the proposed extended semi-parametric AFT cure model provided more efficient and less biased estimations than traditional semi-parametric AFT cure model; higher efficiency and smaller bias were associated with higher sensitivity and specificity of the diagnostic procedures. The proposed method was illustrated using a clinical data example.

Keywords: Cure Model, Expectation-Maximization (EM) Algorithm, Accelerated Failure Time (AFT), Relative Efficiency, Sensitivity and Specificity

Introduction

A cure model is a useful model for analyzing failure time data when there is evidence of long-term survivors. It is assumed that in traditional cure models the cured and uncured status in the censored set cannot be distinguished. However, in many practices, some diagnostic procedures may provide partial information about the cured or uncured status with some sensitivity and specificity. Traditional cure models do not take advantage of this additional information. Recently, Wu *et al*. (2014a; 2014b) proposed a method, called the extended cure model, which incorporated such additional diagnostic cured status information into the traditional Proportional Hazards (PH) cure model analysis. However, in many applications, semi-parametric AFT cure models may be of interest itself and/or may be more appropriate because it does not need the PH assumption and can directly model time to event instead of hazard. In this work, we extended the method of Wu *et al*. (2014a) to AFT cure models.

For traditional cure models, both the Cox PH and AFT cure models have been extensively studied. Let T denote a non-negative random variable for the failure time of interest, x and z the covariate vectors, $\pi(z)$ the individual's uncured probability depending on z and $f(t|x, z)$ and $S(t|x, z)$ the density and the survival function for T, respectively. Assume that $f_u(t|x)$ and $S_u(t|x)$ are the uncured individual's

probability density function (pdf) and the survival function depending on x. We can express the mixture cure model as $f(t|x, z) = \pi(z) f_u(t|x)$ or:

$$S(t \mid x,z) = \pi(z) S_u(t \mid x) + \left[1 - \pi(z)\right]. \quad (1)$$

Logistic regression is commonly used to model the "incidence" part $\pi(z)$, although other links or non-linear regression models can also be used. Parametric, semi-parametric, or non-parametric models could be used for the "latency" part $S_u(t|x)$. The parametric approach includes the following commonly used distributions: Exponential (Jones *et al*., 1981; Ghitany and Maller, 1992), Weibull (Farewell, 1982; 1986), Lognormal (Boag, 1949; Gamel *et al*., 1990), Gompertz (Cantor and Shuster, 1992; Gordon, 1990a; 1990b), Extended generalized gamma (Yamaguchi, 1992) and Generalized F distributions (Peng *et al*., 1998). The non-parametric approach, Kaplan-Meier estimation method, is used without adjusting for the covariate vector x as done in Taylor (1995). The semi-parametric approach includes the Cox PH model (Kuk and Chen, 1992; Peng and Dear, 2000; Sy and Taylor, 2000) and semi-parametric AFT models (Li and Taylor, 2002; Zhang and Peng, 2007). Although a parametric cure model can achieve the greatest efficiency in estimation when its distributional assumption is satisfied, in practice, it can be challenging to justify the assumption. A semi-parametric model does not require a distributional assumption, but may lose efficiency in estimation, compared to a parametric model when a distribution can be correctly identified.

In this work, our main focus is on the evaluation of the performance of the proposed extended semiparametric AFT cure model that incorporates the additional diagnostic information. We performed extensive simulations and demonstrated that, compared to the traditional AFT model, the extension provided more efficient and less biased estimations and higher efficiency and smaller bias were associated with higher sensitivity and specificity of the diagnostic procedures. Finally, we applied the extended semi-parametric AFT cure model to a data example from a pediatric bone fracture study where the Kaplan-Meier curves show that there is a clear cure indication in this dataset (Fig. 1A), suggesting the appropriateness to use a cure model for the analysis. The application of the proposed method showed that the efficiency gain may change the significance (p-values) of some effects after the additional cure information was incorporated. This paper was organized as follows. In Section 2, we extended the traditional semi-parametric AFT cure models to incorporate the additional cure information. In Section 3, the extended cure models were evaluated through extensive simulation studies. In Section 4, we illustrated the use of proposed extended models by a data example from a pediatric bone fracture study. Discussion was given in Section 5.

Accelerated Failure Time Cure Models with Sensitivity and Specificity

Model Specification

Let $\{(t_i, \delta_i, x_i, z_i), i = 1,2,...n\}$ be a data set. Here t_i denotes the i^{th} patient's observed survival time. δ_i is the censoring indicator, which is 0 if t_i is censored and 1 if uncensored (i.e., observed). x_i and z_i are two covariate vectors. Assume that β and γ are the parameter vectors for x_i and z_i, respectively. If the data set is modeled by the semi-parametric AFT cure model specified in (1):

$$\begin{aligned} &\pi(z_i) = \exp(\gamma' z_i) / \left[1 + \exp(\gamma' z_i)\right], \\ &\log(t_i) = \beta' x_i + \varepsilon_i, \end{aligned} \quad (1)$$

where, the error term ε_i has the pdf f_ε and survival function S_ε that have no particular parametric forms. It is noted that $f_u(t_i|x_i) = f_\varepsilon(\log(t_i) - \beta' x_i)/t_i$ and $S_u(t_i|x_i) = S_\varepsilon(\log(t_i) - \beta' x_i)$. Assume $O_0 = \{(t_i, \delta_i, x_i, z_i), i = 1,2,...,n\}$ and $\theta_0' = (\beta', \gamma')$. We can express the observed likelihood as:

$$\begin{aligned} L_o(\theta_0; O_0) = &\prod_{i=1}^{n} \left[\pi(z_i) f_u(t_i \mid x_i)\right]^{\delta_i} \\ &\times \left\{\pi(z_i) S_u(t_i \mid x_i) + \left[1 - \pi(z_i)\right]\right\}^{1-\delta_i}. \end{aligned} \quad (2)$$

Assume that for censored patients, the result d_i from a diagnostic procedure is also observed, which is 1 if patient i is diagnosed as cured; 0 otherwise. A diagnostic procedure usually is associated with certain sensitivity and specificity. Sensitivity measures the proportion of actual positives that are identified correctly (e.g., the percentage of sick people identified correctly as sick). Specificity measures the proportion of negatives that are identified correctly (e.g., the percentage of healthy people identified correctly as healthy). Suppose the diagnostic procedure result does not depend on time, i.e., d_i does not depend on t_i. Assume the diagnostic procedure has a specificity of $1-p_1$ and a sensitivity of p_0. We will have $p_0 \geq p_1$ for a validated diagnostic procedure. Although p_0 and p_1 might be modeled, they are assumed to be independent of any covariates for simplicity. Let $\theta' = (\theta_0', p_0, p_1)$ and $O_1 = \{(t_i, \delta_i, x_i, z_i, d_i), i = 1,2,...,n\}$. For uncensored patients ($\delta_i = 1$), the contribution to the likelihood is the same as that in (2); while for censored patients ($\delta_i = 0$), with the independence assumption of d_i and t_i, the contribution is $p_0^{d_i}(1-p_0)^{1-d_i}\left[1-\pi(z_i)\right]$ if they are cured and the contribution is $p_0^{d_i}(1-p_1)^{1-d_i}\pi(z_i)S_u(t_i|x_i)$ if they are uncured, so the observed likelihood is:

$$L_o\left(\theta;O_1\right)=\prod_{i=1}^{n}\left[\pi\left(z_i\right)f_u\left(t_i\mid x_i\right)\right]^{\delta_i}\begin{Bmatrix}p_1^{d_i}\left(1-p_1\right)^{1-d_i}\pi\left(z_i\right)S_u\left(t_i\mid x_i\right)\\+p_0^{d_i}\left(1-p_0\right)^{1-d_i}\left[1-\pi\left(z_i\right)\right]\end{Bmatrix}^{1-\delta_i}. \tag{3}$$

The diagnostic procedure results are not available for all the censored subjects, so let $\eta_i = 1$ denote the i^{th} subject's diagnostic result available and $\eta_i = 0$ her/his result unavailable. Thus we can write the observed likelihood as:

$$\begin{aligned}L_o\left(\theta;O\right)&=\prod_{i=1}^{n}\left[\pi\left(z_i\right)f_u\left(t_i\mid x_i\right)\right]^{\delta_i}\\&\times\begin{bmatrix}p_1^{d_i}\left(1-p_1\right)^{1-d_i}\pi\left(z_i\right)S_u\left(t_i\mid x_i\right)\\+p_0^{d_i}\left(1-p_0\right)^{1-d_i}\left(1-\pi\left(z_i\right)\right)\end{bmatrix}^{\left(1-\delta_i\right)\eta_i}\\&\times\left[\pi\left(z_i\right)S_u\left(t_i\mid x_i\right)+\left(1-\pi\left(z_i\right)\right)\right]^{\left(1-\delta_i\right)\left(1-\eta_i\right)},\end{aligned} \tag{4}$$

where, $O = \{(t_i, \delta_i, x_i, z_i, \eta_i, d_i), i = 1,2,...,n\}$. It is noted that Equation (4) reduces to Equation (2) except for a constant multiplier when $p_0 = p_1$, which means that if both (1-specificity) and sensitivity are the same, the likelihood functions without and with the diagnostic information are the same. In practice, we want both sensitivity and specificity to be high and $p_0 \neq p_1$. The "incidence" part $\pi(z)$ of the mixture model is modeled by logistic regression. The "latency" part $f_u(t|x)$ and $S_u(t|x)$ of the mixture model is modeled by the semi-parametric AFT model. To estimate the extended semi-parametric AFT cure model parameters in (4), we adapt the Expectation-Maximization (EM) algorithm of Peng (2003) to our extended AFT model. Details of the EM procedure are provided in the following section. To implement the EM algorithm and obtain the parameter estimates, one can apply the method of Li and Taylor (2002) or Zhang and Peng (2007), as shown in the simulation studies.

EM Algorithm for Estimation of the Extended Semi-Parametric AFT Cure Model

Assume that c_i is the indicator of the i^{th} patient's cured status, which is 1 if s/he is not cured (susceptible) and 0 if s/he is cured, where although c_i is a latent variable, because $\delta_i = 1$ implies $c_i = 1$, it is partially observed. The conditional distributions of d_i are as follows:

$$\begin{aligned}d_i\mid\left(c_i=0,\delta_i=0\right)&\sim Bernoulli\left(p_0\right),\\d_i\mid\left(c_i=1,\delta_i=0\right)&\sim Bernoulli\left(p_1\right).\end{aligned}$$

Let $c = \{c_i, i = 1,...,n\}$. The complete log-likelihood can be written as:

$$\begin{aligned}&\ell_c\left(\theta;O,c\right)=\log L_c\left(\theta;O,c\right)\\&=\sum_{i=1}^{n}\left\{c_i\log\left(\pi\left(z_i\right)\right)+1\left(1-c_i\right)\left(1-\delta_i\right)\log\left[1-\pi\left(z_i\right)\right]\right\}\\&+\sum_{i=1}^{n}\left\{c_i\delta_i\log\left(f_u\left(t_i\mid x_i\right)\right)+c_i\left(1-\delta_i\right)\log\left[S_u\left(t_i\mid x_i\right)\right]\right\}\\&+\sum_{i=1}^{n}\left[d_i\log p_1+\left(1-d_i\right)\log\left(1-p_1\right)\right]c_i\left(1-\delta_i\right)\eta_i\\&+\sum_{i=1}^{n}\left[d_i\log p_0+\left(1-d_i\right)\log\left(1-p_0\right)\right]\left(1-c_i\right)\left(1-\delta_i\right)\eta_i,\end{aligned} \tag{5}$$

where:

$$\begin{aligned}f_u\left(t_i\mid x_i\right)&=\frac{f_\varepsilon\left(\log\left(t_i\right)-\beta' x_i\right)}{t_i}\,and\,S_u\left(t_i\mid x_i\right)\\&=S_\varepsilon\left(\log\left(t_i\right)-\beta' x_i\right),\end{aligned}$$

according to the AFT model. Because of $(1-c_i)(1-\delta_i) = 1-c_i$ and $c_i\delta_i = \delta_i$, one can further simplify $\ell_c(\theta; O, c)$ in (5) to:

$$\begin{aligned}&\ell_c\left(\theta;O,c\right)=\sum_{i=1}^{n}\left\{c_i\log\left[\pi\left(z_i\right)\right]+\left(1-c_i\right)\log\left[1-\pi\left(z_i\right)\right]\right\}\\&+\sum_{i=1}^{n}\left\{\delta_i\log\left[h_u\left(t_i\mid x_i\right)\right]+c_i\log\left[S_u\left(t_i\mid x_i\right)\right]\right\}\\&+\sum_{i=1}^{n}\left[d_i\log p_1+\left(1-d_i\right)\log\left(1-p_1\right)\right]c_i\left(1-\delta_i\right)\eta_i\\&+\sum_{i=1}^{n}\left[d_i\log p_0+\left(1-d_i\right)\log\left(1-p_0\right)\right]\left(1-c_i\right)\eta_i\\&=\ell_{c1}\left(\gamma;O,c\right)+\ell_{c2}\left(\beta;O,c\right)+\ell_{c3}\left(p_0,p_1;O,c\right).\end{aligned} \tag{6}$$

Here:

$$h_u\left(t_i\mid x_i\right)=\frac{f_u\left(t_i\mid x_i\right)}{S_u\left(t_i\mid x_i\right)}=\frac{f_\varepsilon\left(\log t_i-\beta' x_i\right)}{t_i S_\varepsilon\left(\log t_i-\beta' x_i\right)}$$

is the hazard function of the failure time of uncured patients:

$$\begin{aligned}\ell_{c1}\left(\gamma;O,c\right)&=\sum_{i=1}^{n}\left\{c_i\log\left[\pi\left(z_i\right)\right]+\left(1-c_i\right)\log\left[1-\pi\left(z_i\right)\right]\right\},\\\ell_{c2}\left(\beta;O,c\right)&=\sum_{i=1}^{n}\left\{\delta_i\log\left[h_u\left(t_i\mid x_i\right)\right]+c_i\log\left[S_u\left(t_i\mid x_i\right)\right]\right\},\\\ell_{c3}\left(p_0,p_1;O,c\right)&=\sum_{i=1}^{n}\left[d_i\log p_1+\left(1-d_i\right)\log\left(1-p_1\right)\right]c_i\left(1-\delta_i\right)\eta_i\\&+\sum_{i=1}^{n}\left[d_i\log p_0+\left(1-d_i\right)\log\left(1-p_0\right)\right]\left(1-c_i\right)\eta_i.\end{aligned}$$

Equation (6) shows that the complete log-likelihood function can be separated into three parts: the first part $\ell_{c1}(\gamma,\ O,\ c)$ contains only the "incidence" parameter vector γ for the covariate vector z, the second part $\ell_{c2}(\gamma,\ O,\ c)$ contains only the "latency" parameter vector β for the covariate vector x and the third part $\ell_{c3}(p_0,\ p_1,\ O,\ c)$ contains only the specificity parameter $1\text{-}p_1$ and sensitivity parameter p_0. Therefore, we can maximize separately the three parts given c and carry out the EM algorithm in the following steps.

Initial value: Let $\theta^{(0)}$ be an initial value to start the EM algorithm.

E-step: The E-step is to calculate the expectation of the complete log-likelihood function $\ell_c(\theta)$, conditional on the observed data and $\theta^{(r)}$ the estimate of θ at the r^{th} iteration. That is, calculate the following conditional expectation:

$$w_i^{(r)} = E\left(c_i \mid \theta^{(r)}, O\right) = P\left(c_i = 1 \mid \theta^{(r)}, O\right),$$

which is the estimate of the i^{th} patient's uncured probability at the r^{th} iteration. Because:

$$\begin{aligned}
&P\left(c_i = 1 \mid d_i = 1, \delta_i = 0, \theta^{(r)}, O\right)\\
&= \eta_i \frac{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right) p_1^{(r)}}{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right) p_1^{(r)} + \left[1-\pi^{(r)}\left(z_i\right)\right] p_0^{(r)}}\\
&+\left(1-\eta_i\right)\frac{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right)}{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right) + \left[1-\pi^{(r)}\left(z_i\right)\right]}
\end{aligned}$$

and:

$$\begin{aligned}
&P\left(c_i = 1 \mid d_i = 1, \delta_i = 0, \theta^{(r)}, O\right)\\
&= \eta_i \frac{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right)\left(1-p_1^{(r)}\right)}{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right)\left(1-p_1^{(r)}\right) + \left[1-\pi^{(r)}\left(z_i\right)\right]\left(1-p_1^{(r)}\right)}\\
&+\left(1-\eta_i\right)\frac{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right)}{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right) + \left[1-\pi^{(r)}\left(z_i\right)\right]},
\end{aligned}$$

one can express $w_i^{(r)}$ as:

$$\begin{aligned}
&w_i^{(r)} = P\left(c_i = 1 \mid \theta^{(r)}, O\right)\\
&= \delta_i + \left(1-\delta_i\right) d_i \eta_i \frac{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right) p_1^{(r)}}{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right) p_1^{(r)} + \left[1-\pi^{(r)}\left(z_i\right)\right] p_0^{(r)}}\\
&+\left(1-\delta_i\right)\left(1-d_i\right)\eta_i \frac{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right)\left(1-p_1^{(r)}\right)}{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right)\left(1-p_1^{(r)}\right) + \left[1-\pi^{(r)}\left(z_i\right)\right]\left(1-p_0^{(r)}\right)}\\
&+\left(1-\delta_i\right)\left(1-\eta_i\right)\frac{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right)}{\pi^{(r)}\left(z_i\right) S_u^{(r)}\left(t_i \mid x_i\right) + \left[1-\pi^{(r)}\left(z_i\right)\right]}.
\end{aligned} \tag{7}$$

M-step: The M-step is to maximize the expected complete log-likelihood function with respect to θ to obtain $\theta^{(r+1)}$, which is the sum of the following three functions:

$$\begin{aligned}
&\tilde{\ell}_{c1}\left(\gamma \mid w^{(r)}, O\right)\\
&= \sum_{i=1}^{n}\left\{w_i^{(r)} \log\left(\pi\left(z_i\right)\right) + \left(1-w_i^{(r)}\right)\log\left[1-\pi\left(z_i\right)\right]\right\},
\end{aligned} \tag{8}$$

$$\begin{aligned}
&\tilde{\ell}_{c2}\left(\beta \mid w^{(r)}, O\right)\\
&= \sum_{i=1}^{n}\left\{\delta_i \log\left[h_u\left(t_i \mid x_i\right)\right] + w_i^{(r)} \log\left[S_u\left(t_i \mid x_i\right)\right]\right\},
\end{aligned} \tag{9}$$

$$\begin{aligned}
&\tilde{\ell}_{c3}\left(p_0, p_1 \mid w^{(r)}, O\right)\\
&= \sum_{i=1}^{n}\left[d_i \log\left(p_1\right) + \left(1-d_i\right)\log\left(1-p_1\right)\right] w_i^{(r)}\left(1-\delta_i\right)\eta_i\\
&+\sum_{i=1}^{n}\left[d_i \log\left(p_0\right) + \left(1-d_i\right)\log\left(1-p_0\right)\right]\left(1-w_i^{(r)}\right)\eta_i
\end{aligned} \tag{10}$$

for $w^{(r)} = \left\{w_i^{(r)}, i = 1,...,n\right\}$.

Because Equation 8 is the log-likelihood function of a logistic regression model for values arising from a Bernoulli distribution with the response probability $\pi(z_i) = \exp(\gamma' z_i) / [1+ \exp(\gamma' z_i)]$, the usual optimization methods such as the Newton-Raphson method can be used to maximize this log-likelihood function, which can be carried out in most standard logistic regression packages to obtain the estimate of γ.

For Equation 9, the maximization of $\tilde{\ell}_{c2}\left(\beta \mid w^{(r)}, O\right)$ involves the joint estimation of β and S_ε. This maximization can obtain the estimates $\beta^{(r+1)}$ and $S_\varepsilon^{(r+1)}$ e by using the grid search method of Li and Taylor (2002). One can also use the linear programming approach to obtain $\beta^{(r+1)}$ first by minimizing the gradient of a convex function. After $\beta^{(r+1)}$ is obtained, $S_\varepsilon^{(r+1)}$ can be estimated based on the residuals as done in Zhang and Peng (2007).

For Equation (10), $p_0^{(r+1)}$ and $p_1^{(r+1)}$ can be obtained explicitly. The following is the updating formula:

$$\begin{aligned}
p_0^{(r+1)} &= \frac{\sum_{i=1}^{n} d_i \eta_i \left(1-w_i^{(r)}\right)\left(1-\delta_i\right)}{\sum_{i=1}^{n} \eta_i \left(1-w_i^{(r)}\right)\left(1-\delta_i\right)} = \frac{\sum_{i:\delta_i=0\ \&\ \eta_i=1} d_i \left(1-w_i^{(r)}\right)}{\sum_{i:\delta_i=0\ \&\ \eta_i=1}\left(1-w_i^{(r)}\right)},\\
p_1^{(r+1)} &= \frac{\sum_{i=1}^{n} d_i \eta_i w_i^{(r)}\left(1-\delta_i\right)}{\sum_{i=1}^{n} \eta_i w_i^{(r)}\left(1-\delta_i\right)} = \frac{\sum_{i:\delta_i=0\ \&\ \eta_i=1} d_i w_i^{(r)}}{\sum_{i:\delta_i=0\ \&\ \eta_i=1} w_i^{(r)}}.
\end{aligned}$$

It is noted that there is no need to estimate p_0 and p_1 if the sensitivity and specificity are known externally from the diagnostic procedure.

Iteration: The algorithm is iterated until $||\theta^{(r+1)}-\theta^{(r)}||$ is sufficiently small.

Although the method of Louis (1982) may be used to estimate the variance of the EM estimators, we follow Peng's suggestion (Peng, 2003) to use the bootstrap method to estimate the variance of the estimated parameters.

Evaluations of the Extended Semi-Parametric AFT Cure Model

Simulation Setup

To assess the performance of the extended semi-parametric AFT cure model, we compared the following three models through extensive simulations: (i) The traditional semi-parametric AFT cure model without diagnostic information, (ii) the extended semi-parametric AFT cure model incorporating diagnostic information with unknown sensitivity and specificity and (iii) the extended semi-parametric AFT cure model incorporating diagnostic information with sensitivity and specificity known a priori. In estimating model parameters, we adapted the approaches of Li and Taylor (2002) (LT) and Zhang and Peng (2007) (ZP) and compared the performances between these two methods.

To mimic the pediatric bone data, we first generated c_i according to the incidence model with evenly distributed three-level covariate TRT and two-level covariate SEX:

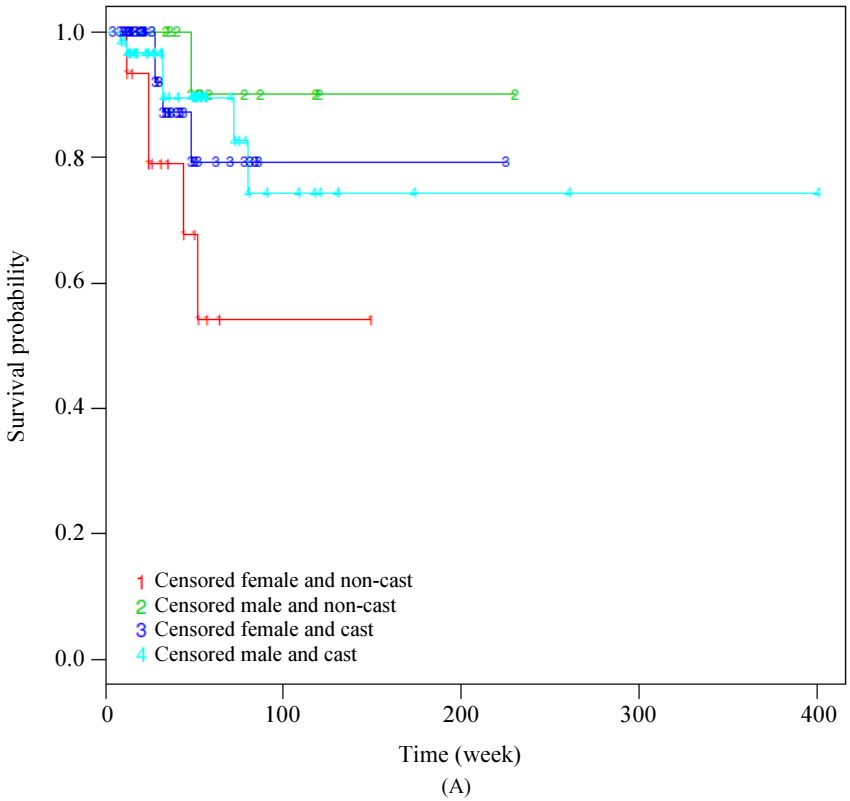
$$\log it\left(\pi_i\right) = \gamma_0 + \gamma_1 I_{(TRT_i=1)} + \gamma_2 I_{(TRT_i=2)} + \gamma_3 I_{(SEX_i=Male)}. \quad (11)$$

Here π_i is the i^{th} subject's uncured probability. The true parameter values were $\gamma_0 = 0.25$, $\gamma_1 = -0.1$, $\gamma_2 = 0.5$ and $\gamma_3 = -0.1$. Survival data were simulated for the latency part, according to the Weibull AFT model:

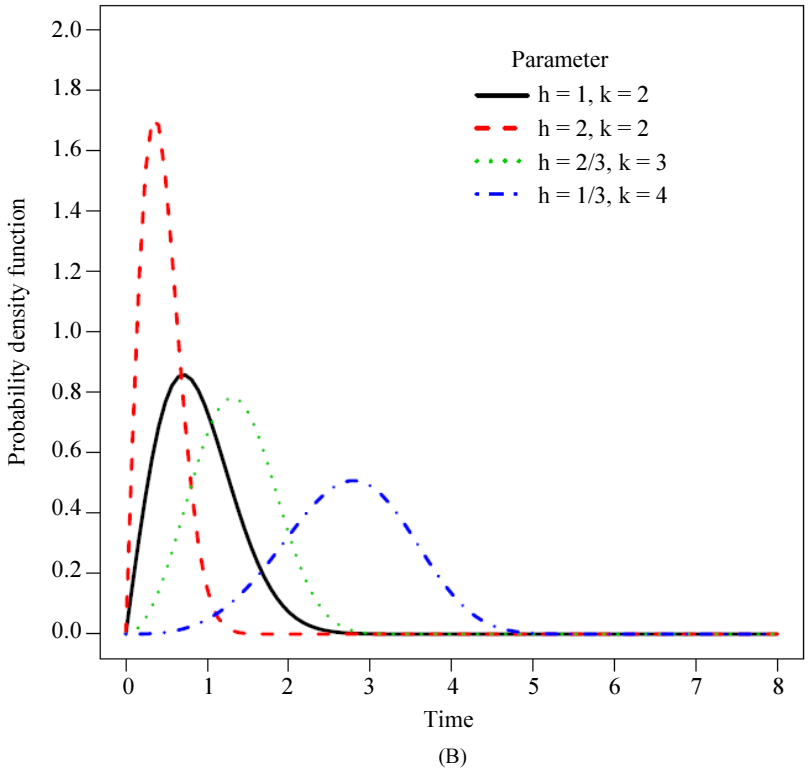
$$\log\left(t_i\right) = \beta_1 I_{(TRT_i=1)} + \beta_2 I_{(TRT_i=2)} + \beta_3 I_{(SEX_i=Male)} + \varepsilon_i, \quad (12)$$

with the true parameter values as $\beta_1 = 0.2$, $\beta_2 = -0.3$, $\beta_3 = 0.1$ and the baseline survival function as $S_0(t|k, h) = \exp[-(ht)^k]$. Four different sets of shape and scale parameters (h,k) were considered:

$$(1)h=1,k=2;(2)h=2,k=2;$$
$$(3)h=\frac{2}{3},k=3;(4)h=\frac{1}{3},k=4, \quad (13)$$



(A)



(B)

Fig. 1: K-M curve for PPC and baseline Weibull probability density functions; (A). Kaplan-Meier (K-M) curve for time to PPC by treatment and gender; (B). Curves of baseline Weibull probability density function $f_0(t|k,h) = h_0(t|k,h)\exp[-\int_0^t h_0(u|k,h)du] = kh(ht)^{k-1}\exp[-(ht)^k]$

(see these four baseline Weibull pdf shapes in Fig. 1B). We set the maximum survival time to 6 and simulated the censoring time from uniform (0,6). As a result, the expected censoring rate of the four parameter sets was 14.8% when $h = 1$, $k = 2$; 7.4% when $h = 2$, $k = 2$; 22.3% when $h = \frac{2}{3}$, k = 3; 45.3% when $h = \frac{1}{3}$, $k = 4$. After c_i and the censoring status were determined, d_i was simulated from the following Bernoulli distributions:

$$d_i \mid \left(c_i = 0, \delta_i = 0\right) \sim Bernoulli\left(p_0\right),$$
$$d_i \mid \left(c_i = 1, \delta_i = 0\right) \sim Bernoulli\left(p_1\right).$$

True sensitivity of 70% and 100% and true specificity (1-p_1) of 100% were used. Simulations with 100% subjects having available diagnostic information were performed for all settings described above. For each simulation configuration, 200 subjects were simulated and a total of 1,000 simulation runs were performed. In each simulation run, the variances of estimated parameters were based on 1,000 bootstrap samples.

Simulation Results

Figures 2 to 4 show the performance of regression parameter estimates in terms of bias, Mean Squared Error (MSE) and Relative Efficiency (RE) to the traditional semi-parametric AFT cure model, respectively. From the top to the bottom in each figure, the odd rows are for the models with known sensitivity and specificity, while the even rows are the models with unknown sensitivity and specificity. Numbers 1 to 4 correspond to the four different parameter combinations of the baseline Weibull distribution in Equation (13). Subscripts "LT" and "ZP" indicate the use of LT and ZP estimation methods, respectively, in fitting the extended semi-parametric AFT cure model incorporating diagnostic information with known sensitivity and specificity. Subscripts of "LTu" and "ZPu" are for the estimation methods of the same model incorporating diagnostic information with unknown sensitivity and specificity. In using the LT approach, we applied the non-linear minimization method to obtain the parameter estimates, implemented by nlm function in R, instead of the grid search approach in their original method.

LT Estimation Method for the Extended Semi-Parametric AFT Cure Model

Simulation results of the extended semi-parametric AFT cure model compared to the traditional method ("CL"), all adapting the LT method, are presented in the

first two rows of Figures 2 to 4. Notice that results for (h, k) = $\left(\frac{1}{3},4\right)$ are not shown because of excessively large bias and MSE. Further investigation of this case is shown in Table 1 and discussed later. In Figure 2, the bias of the latency parameter estimates is large, likely due to the use of non-monotonic estimation function in Equation (12) of Li and Taylor (2002). RE gain slightly increases with censoring rate (e.g., '1', '2' Vs. '3' in Figure 4). RE gain is more when sensitivity and specificity are known than that when they are unknown.

Because the baseline Weibull distribution with $\left(h=\frac{1}{3},k=4\right)$ has a high censoring rate of 45%, a possible explanation for the large bias, especially in the latency part, may be due to the use of the zero-tail completion in estimating the survival functions. In Table 1, we explored the effect of exponential-tail completion for the case when the baseline Weibull distribution $\left(h=\frac{1}{3},k=4\right)$ is used. Compared to the zero-tail completion, the exponential-tail completion does not improve much.

ZP Estimation Method for the Extended Semi-Parametric AFT Cure Model

Simulation results of the extended semiparametric AFT cure model compared to the traditional method ("CL"), all using the ZP method, are presented in the third and fourth rows of Figures 2 to 4 for unknown and known sensitivity and specificity, respectively. Instead of using linear programming suggested by Zhang and Peng (2007), the non-linear minimization is implemented by nlm function in R to search for parameter estimates.

In general, the bias and MSE of the extended model is smaller than those of the traditional method. RE of the extended model is increased with sensitivity and censoring rate. Moreover, the RE curves of the γ_1 estimate for parameter set 4 $\left(h=\frac{1}{3},k=4\right)$ are not shown because of large value (> 3). The gain in RE and reduction in MSE and bias are larger with known sensitivity and specificity than with unknown sensitivity and specificity. The improvement is quite significant especially for parameter set 4 $\left(h=\frac{1}{3},k=4\right)$.

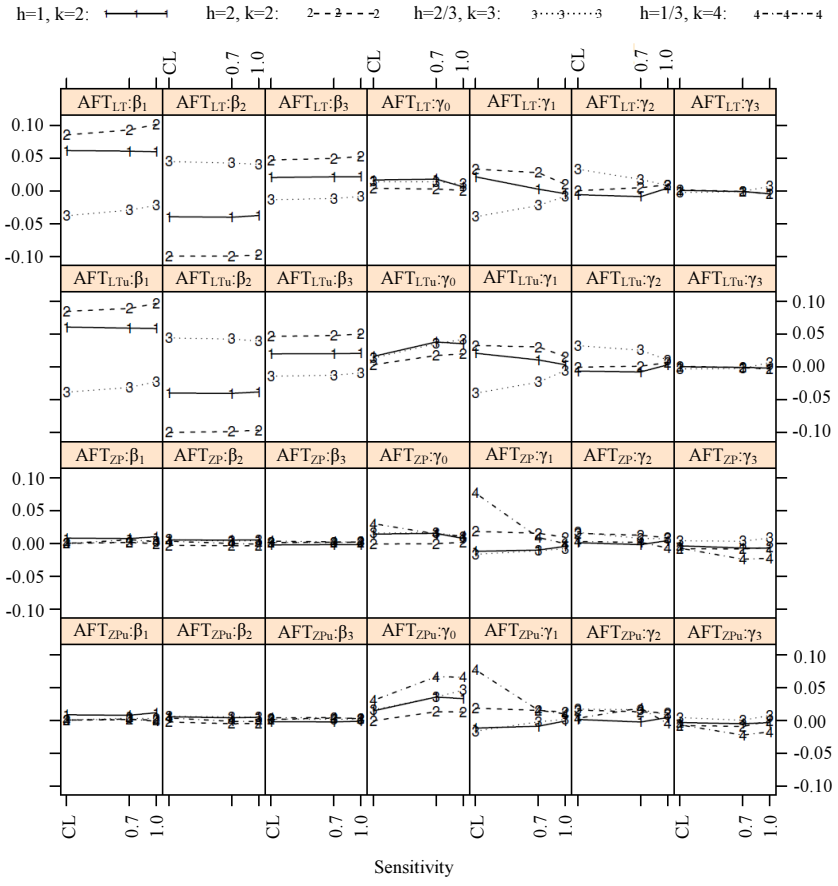


Fig. 2: Bias

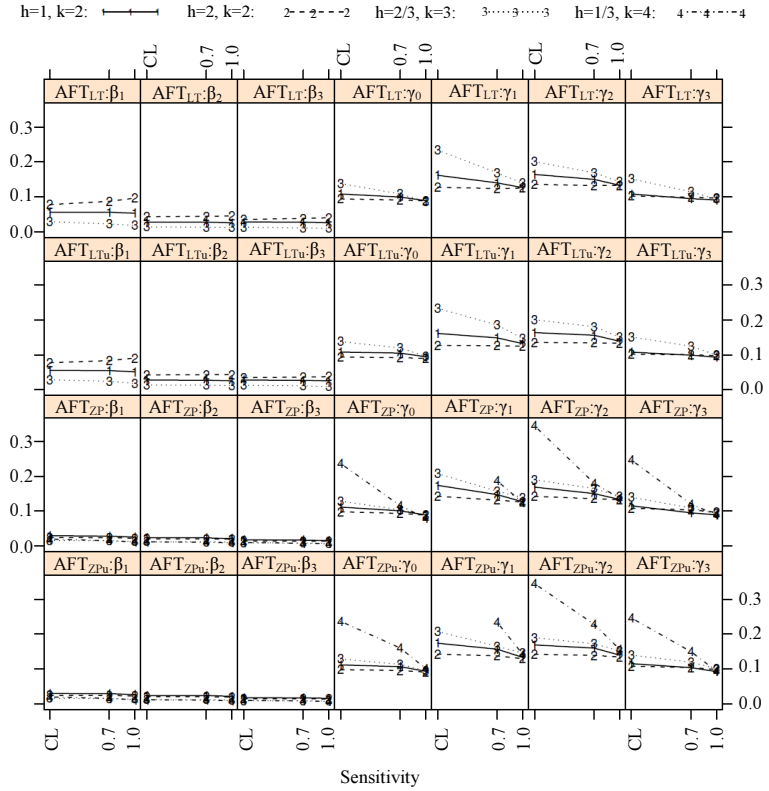


Fig. 3: Mean squared error (MSE)

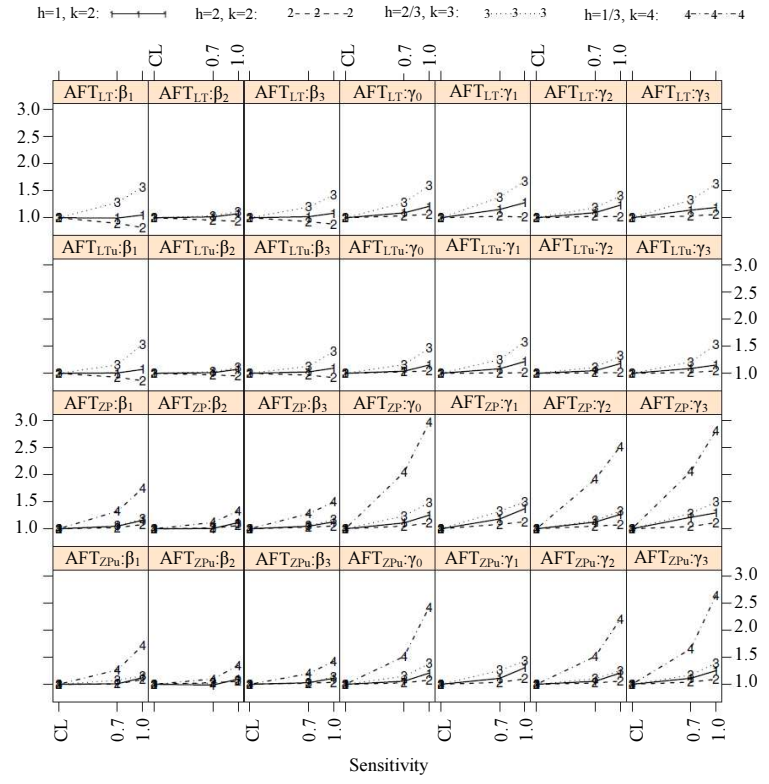


Fig. 4: Relative efficiency

Table 1: Simulation results for models with baseline Weibull hazard $\left(h=\frac{1}{3},k=4\right)$ - assume known sensitivity and specificity, zero- and exponential-tail completions (Using Li and Taylor's estimation method)

				p_0 for Extended Model			
				Zero Tail Completion		Exponential Tail Completion	
	Statistics	True Parameter	Classic Model	0.7	1	0.7	1
β_1	Mean	0.2	-3.91	-0.86	0.05	-0.86	0.05
	Bias		-4.11	-1.06	-0.15	-1.06	-0.15
	SD		9.66	3.08	0.21	3.05	0.26
	MSE		110.19	10.60	0.07	10.42	0.09
β_2	Mean	-0.3	0.16	-0.05	-0.09	-0.06	-0.08
	Bias		0.46	0.25	0.21	0.24	0.22
	SD		3.37	0.38	0.14	0.44	0.16
	MSE		11.60	0.21	0.06	0.25	0.075
β_2	Mean	0.1	-1.58	-0.23	0.02	-0.24	0.02
	Bias		-1.68	-0.33	-0.08	-0.34	-0.08
	SD		17.67	0.98	0.15	1.19	0.17
	MSE		315.22	1.06	0.03	1.52	0.04
γ_0	Mean	0.25	2.06	0.34	0.26	0.36	0.26
	Bias		1.81	0.09	0.01	0.11	0.01
	SD		5.09	0.48	0.29	0.47	0.29
	MSE		29.17	0.24	0.08	0.24	0.08
γ_1	Mean	-0.1	-1.45	-0.46	-0.12	-0.47	-0.12
	Bias		-1.35	-0.36	-0.02	-0.37	-0.02
	SD		6.09	0.82	0.37	0.81	0.37
	MSE		38.94	0.80	0.14	0.80	0.14
γ_2	Mean	0.5	1.59	0.64	0.50	0.63	0.49
	Bias		1.09	0.14	-0.002	0.13	-0.01
	SD		6.68	0.54	0.37	0.53	0.37
	MSE		45.83	0.30	0.14	0.30	0.14
γ_3	Mean	-0.1	-0.59	-0.23	-0.13	-0.24	-0.13
	Bias		-0.49	-0.13	-0.03	-0.14	-0.03
	SD		3.46	0.52	0.31	0.52	0.31
	MSE		12.22	0.29	0.10	0.29	0.10

Comparison of ZP and LT Methods for Estimation of the Extended Semi-Parametric AFT Cure Model

To compare the simulation results between the estimation methods of ZP and LT, we first consider the bias. Overall, the point estimates are consistent for the ZP method, while the LT method is more likely to produce non-consistent estimates, especially when the baseline Weibull distribution with parameter set 4 $\left(h=\frac{1}{3},k=4\right)$, which produced highest censoring rate, is used.

As for the RE gains in the parameter sets 1 to 3, the ZP and LT methods are similar. Parameter set 4 is not compared because of the large bias using LT method.

In conclusion, based on our simulation results, the ZP method provides better estimations than the LT method for the extended semi-parametric AFT cure model.

Real Example: Pediatric Bone Data

This was a retrospective clinical study that 157 (75 girls and 82 boys) children's charts were reviewed to identify the incidence of premature physeal closure (PPC) following physeal fractures of distal end of tibia (Leary *et al*., 2009). Sixteen out of these 157 children were identified as having PPC. Children were considered cured if the symmetric Harris growth arrest line was observed or closure of the growth plate was seen radiographically. As a result, ninety-six children were considered cured. Because the remaining 45 children's diagnostic cured statuses could not be determined, their diagnostic cured statuses were considered unavailable.

As shown in the Kaplan-Meier curve of the time to PPC (Figure 1A), there is a clear cure indication in this data set. The semi-parametric AFT cure model was used for the data analysis as an illustration. The ascertainment of cure using the symmetric Harris growth arrest line or closure of the growth plate was considered definitive, so it was treated as a diagnostic procedure with known 100% sensitivity and specificity. We included the factor of treatment methods (Cast and non-Cast) and gender in the survival portion and the cure portion of the semi-parametric AFT cure model.

Table 2: Comparisons of applications of semi-parametric Accelerated Failure Time (AFT) cure model without and with diagnostic information to pediatric bone data

	Traditional cure model (without diagnostic information)			Extended cure model (with diagnostic information)		
Survival Portion:						
Effect	log(TR*)	SE	p-value	log(TR)	SE	p-value
Male	-0.000	0.784	>0.999	0.134	0.509	0.793
Cast	-0.223	0.731	0.760	0.151	0.453	0.739
Logistic Portion:						
Effect	log(OR)	SE	p-value	log(OR)	SE	p-value
Intercept	-0.650	1.834	**0.723**	-1.189	0.508	**0.020**
Male	-0.718	3.791	0.850	-0.344	0.475	0.469
Cast	-0.074	4.382	**0.987**	-0.896	0.510	**0.079**

*: Ratio of survival times

Table 2 shows the analysis result of fitting the semi-parametric AFT cure model with and without the diagnostic information included to the pediatric bone data. The ZP method was used to estimate the semiparametric, traditional and the extended semiparametric AFT cure model parameters. During the bootstrapping step, if the point estimate had an absolute value over 1,000, the bootstrap sample was treated as not converged. The estimates of the parameters in the survival portion showed different signs, while the pvalues suggested non-significant conclusions. The standard errors from the extended semi-parametric AFT cure model were much smaller. The 2-sided p-value for logistic intercept was 0.723 in the traditional model and this changed to a significant p-value of 0.020 in the extended model. Notice that the p-value for the Cast factor in the logistic portion was also much smaller in the extended model. These comparisons between the traditional and extended models were consistent with the findings in the simulation results.

Discussion

Othus *et al*. (2012) advocated cure models for analyzing survival data when there is evidence of long-term survivors. It is assumed that in traditional cure models the cured or uncured status in the censored set cannot be distinguished. However, in many studies, there are diagnostic procedures available to provide further information about whether a subject is cured. Wu *et al*. (2014a) proposed a method, called the extended PH cure model, which incorporated such additional diagnostic cured status information into the traditional cure model analysis. In this work, we extended their approach to semi-parametric AFT cure models because the AFT model does not need the PH assumption and can directly model time to event instead of hazard. In this work, we have demonstrated the implementation of the extended semi-parametric AFT cure model and showed that the extended model has the potential to improve the estimation efficiency of the traditional model.

We performed extensive simulations to evaluate the performance of the extended semi-parametric AFT cure model. The simulations showed that the extended model provided more efficient and less biased estimations when the ZP estimation method was used. In contrast, the LT estimation method performed less satisfactorily. In using the ZP estimation method, a large efficiency gain was noted when the censoring rate was high. This may be because when the censoring rate is high and so is the set of subjects with undetermined cured or uncured status, adding additional diagnostic data can provide more information and improve statistical efficiency. In the data example, fitting the extended semi-parametric AFT cure model to pediatric bone data shows a significant efficiency gain, indicated by smaller standard errors, compared to those from the traditional model.

Conclusion

The proposed extended semi-parametric AFT cure model provides an alternative approach to incorporating additional diagnostic information about cure. Failure to use such data would be wasteful and result in efficiency loss. It is highly recommended that when additional cure information is available it should be incorporated into the model. In addition, when designing and conducting studies, it is useful to devise cure diagnostic procedures to collect additional cure status information.

Acknowledgement

The research of YL, SL and WJS was partially supported by NIH/NCI CCSG Grant PCA 072720C.

Author's Contributions

Yu Wu and Yong Lin: Concept and design, analysis and interpretation, writing and final appval.

Shou-En Lu, Chin-Shang Li and Weichung Joe Shih: Concept and design, interpretation, writing and final approval

Ethics

The authors have declared no conflict of interest.

References

Boag, J.W., 1949. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J. Royal Stat. Society, 11: 15-53.

Cantor, A.B. and J.J. Shuster, 1992. Parametric versus non-parametric methods for estimating cure rates based on censored survival data. Stat. Med., 11: 931-937. DOI: 10.1002/sim.4780110710

Farewell, V.T., 1982. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics, 38: 1041-1046. DOI: 10.2307/2529885

Farewell, V.T., 1986. Mixture models in survival analysis: are they worth the risk? Canad. J. Stat., 14: 257-262. DOI: 10.2307/3314804

Gamel, J.W., I.W. McLean and S.H. Rosenberg, 1990. Proportion cured and mean log survival time as functions of tumor size. Stat. Med., 9: 999-1006. DOI: 10.1002/sim.4780090814

Ghitany, M.E. and R.A. Maller, 1992. Asymptotic results for exponential mixture models with long term survivors. Statistics, 23: 321-336. DOI: 10.1080/02331889208802379

Gordon, N.H., 1990a. Maximum likelihood estimation for mixtures of two gompertz distributions when censoring occurs. Commun. Stat. Simulat. Comput., 19: 733-747. DOI: 10.1080/03610919008812885

Gordon, N.H., 1990b. Application of the theory of finite mixtures for the estimation of cure rates of treated cancer patients. Stat. Med., 9: 397-407. DOI: 10.1002/sim.4780090411

Jones, D.R., R.L. Powles, D. Machin and R.J. Sylvester, 1981. On estimating the proportion of cured patients in clinical studies. Biometrie-Praximetrie, 21: 1-11.

Kuk, A.Y.C. and C. Chen, 1992. A mixture model combining logistic regression with proportional hazards regression. Biometrika, 79: 531-541. DOI: 10.1093/biomet/79.3.531

Leary, J.T., M. Handling, M. Talerico, Y. Lin and J.A. Bowe, 2009. Physeal fractures of the distal tibia: predictive factors of premature physeal closure and growth arrest. J. Pediatric Orthopaed., 29: 356-361. DOI: 10.1097/BPO.0b013e3181a6bfe8

Li, C.S. and J.M.G. Taylor, 2002 A semi-parametric accelerated failure time cure model. Stat. Med., 21: 3235-3247. DOI: 10.1002/sim.1260

Louis, T.A., 1982. Finding the observed information matrix when using the em algorithm. J. Royal Stat. Society, 44: 226-233.

Othus, M., B. Barlogie, M.L. LeBlanc and J.J. Crowley, 2012. Cure models as a useful statistical tool for analyzing survival. Clin. Cancer Res., 18: 3731-3736. DOI: 10.1158/1078-0432.CCR-11-2859

Peng, Y., 2003. Estimating baseline distribution in proportional hazards cure models. Comput. Stat. Data Anal., 42: 187-201. DOI: 10.1016/S0167-9473(02)00158-5

Peng, Y. and K.B.G. Dear, 2000. A nonparametric mixture model for cure rate estimation. Biometrics, 56: 237-243. DOI: 10.1111/j.0006-341X.2000.00237.x

Peng, Y., K.B.G. Dear and J.W. Denham, 1998. A generalized f mixture model for cure rate estimation. Stat. Med., 17: 813-830. DOI: 10.1002/(SICI)1097-0258(19980430)17:8<813::AID-SIM775>3.0.CO;2-#

Sy, J.P. and J.M.G. Taylor, 2000. Estimation in a Cox proportional hazards cure model. Biometrics, 56: 227-236. DOI: 10.1111/j.0006-341X.2000.00227.x

Taylor, J.M.G., 1995. Semi-parametric estimation in failure time mixture models. Biometrics, 51: 899-907. DOI: 10.2307/2532991

Wu, Y., Y. Lin, S.E. Lu, C.S. Li and W.J. Shih, 2014a. Extension of a Cox proportional hazards cure model when cure information is partially known. Biostatistics, 15: 540-554. DOI: 10.1093/biostatistics/kxu002

Wu, Y., Y. Lin, C.S. Li, S.E. Lu and W.J. Shih, 2014b. Asymptotic efficiency of an exponential cure model when cure information is partially known. Int. J. Stat. Probability, 3: 1-17. DOI: 10.5539/ijsp.v3n3p1

Yamaguchi, K., 1992. Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of "permanent employment" in Japan. J. Am. Stat. Assoc., 87: 284-292. DOI: 10.2307/2290258

Zhang, J. and Y. Peng, 2007. A new estimation method for the semiparametric accelerated failure time mixture cure model. Stat. Med., 26: 3157-3171. DOI: 10.1002/sim.2748