Pertanika J. Sci. & Technol. 20 (2): 243 - 255 (2012)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Turnbull versus Kaplan-Meier Estimators of Cure Rate Estimation Using Interval Censored Data

Bader Ahmad Aljawadi*, Mohd Rizam Abu Bakar and Noor Akma

Institute for Mathematical Research, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

ABSTRACT

This study deals with the analysis of the cure rate estimation based on the Bounded Cumulative Hazard (BCH) model using interval censored data, given that the exact distribution of the data set is unknown. Thus, the non-parametric estimation methods are employed by means of the EM algorithm. The Turnbull and Kaplan Meier estimators were proposed to estimate the survival function, even though the Kaplan Meier estimator faces some restrictions in term of interval survival data. A comparison of the cure rate estimation based on the two estimators was done through a simulation study.

Keywords: BCH model, interval censored data, cure fraction, EM algorithm, Turnbull estimator, Kaplan Meier estimator

INTRODUCTION

Cancer is one of the major chronic diseases which cause a notable amount of health administrative costs. Prognosis and possible cure from cancer are important measures of lifetimes which can be assessed by analyzing the survival of cancer patients. In survival data from cancer studies, the term cure may refer to a substance or procedure that changes the lifestyle, or may refer to the state of being healed or cured. The proportion of the individuals with a disease that is cured by a given treatment is called the cure fraction.

Survival models incorporating the cure fraction in the analysis known as the cure rate models are being widely used in analyzing data from cancer clinical trials (Zeng *et al.,* 2006). These models were basically developed to estimate the proportion of the patients who are cured as well as the odds of survival of the patients not cured up to a certain point of time (Andreas *et al.,* 2006).

The first cure rate model was published by Boag in 1949 and this was later developed by Berkson and Gage in 1952. In this model, the probability of survival at any given time

Article history:
Received: 29 January 2010
Accepted: 18 January 2012

Email addresses:
Bader_aljawadi@yahoo.com (Bader Ahmad Aljawadi),
mrizam@science.upm.edu.my (Mohd Rizam Abu Bakar),
noorakma@science.upm.edu.my (Noor Akma Ibrahim)
*Corresponding Author

ISSN: 0128-7680 © 2012 Universiti Putra Malaysia Press.

t equals to the proportion of those who are cured (π) plus those who are not cured (1 – π) but have not died. This model is known as the mixture cure rate model which can be defined mathematically, as follows:

$$S(t) = \pi + (1 - \pi)\, S^{*}(t) \tag{1}$$

where $S(t)$ and $S^{*}(t)$ are the survival functions for the entire population and the uncured patients, respectively.

The mixture model plays an important role in reliability and survival analysis, and it is becoming increasingly popular in analyzing data from clinical trials. In fact, the model has been extensively discussed by several authors including Farewell (1986), Gamel *et al.* (1990), Cantor and Shuster (1992), Kuk and Chen (1992), Peng and Dear (2000), Peng and Carreier (2002), Binbing *et al.* (2004), Abu Bakar *et al.* (2009), and in many more recent studies which have been conducted based on this model, as in Kim *et al.* (2009) who proposed a new mixture model via latent cure rate markers for survival data with a cure fraction. Seppa *et al.* (2010) applied a mixture cure fraction model with random effects to cause-specific survival data of female breast cancer patients. The researchers used two sets of random effects to capture the regional variation in the cure fraction and in the survival of the uncured patients, respectively. Furthermore, Castro *et al.* (2010) described an application of the mixture and bounded cumulative hazard models for location, scale, and shape (GAMLSS) framework to the fitting of long-term survival models. On the other hand, Peng and Taylor (2011) considered the mixture cure model with random effects and proposed several estimation methods based on Gaussian quadrature, rejection sampling, and importance sampling to obtain the maximum likelihood estimates of the model for clustered survival data with a cure fraction. Meanwhile, Xiang *et al.* (2011) proposed a mixture cure modelling procedure for analyzing clustered and interval censored survival time data by incorporating random effects in both the logistic regression and PH regression components.

Although this model appears to be attractive and is widely used in survival analysis, Chen *et al.* (1999) stated that it has some drawbacks which include the following:

- When covariates are involved in the analysis, the mixture model does not have a proportional hazard structure.
- The mixture model yields improper posterior distributions for many types of non-informative improper priors when covariates are included through the parameter π via a standard regression model.
- This model does not appear to describe the underlying biological process generating the failure time, at least in the context of cancer relapse.

However, Chen *et al.* (1999) proposed the bounded cumulative hazard (BCH) model developed by Yakovlev *et al.* (1993) as a viable alternative to the mixture model. This model can be derived based on the assumption that for a group of cancer patients entering a clinical trial and after the initial treatment, a number of cancer cells left active and may grow rapidly to produce a detectable cancer mass later on (i.e. cancer relapse). The number of cancer cells denoted by N is assumed to follow Bernoulli, negative binomial or Poisson distribution, whereby considering a Bernoulli distribution is related to the classical mixture model specified in equation (1), where $\pi = P\ (N = 0)$, while considering a negative binomial distribution with parameters α and θ, at the same time, where $\theta = E(N)$ and α are real numbers. For $\theta > 0$ and $\alpha\theta > -1$, the survival function is defined as follows (Rodrigues *et al*., 2009):

$$S(t) = [1 + \alpha\theta F(t)]^{-1/\alpha}$$

where $F(t)$ is the cumulative distribution function.

When N is assumed to follow Poisson distribution with mean θ which is considered as the most attractive assumption since it provides more flexible model (Rodrigues *et al*., 2009). Then, the survival function for the BCH model under this assumption can be obtained by:

$$S(t) = \exp(-\ \theta F(t)) \tag{2}$$

where $F(t)$ is the cumulative distribution function such that $F(t) = 1 - S(t)$. See Chen *et al.* (1999) and Aljawadi *et al*. (2011).

Based on the BCH model defined in (2), the cure fraction (π) can therefore be obtained using:

$$\pi = \lim_{t\to\infty} S(t) = P(N = 0)$$
$$\lim_{t\to\infty} \exp(-\theta F(t)) = \exp(-\theta)$$

Note that when $\theta \to \infty$, then $\pi \to 0$, whereas $\theta \to 0$ then $\pi \to 1$.

In the survival data analysis, the lifetime t can be considered as an exact or censored lifetime; however, other cases often occur in cancer studies, where the follow-up of the patients is a pre-fixed time period or visited periodically for a fixed number of times. In this article, the lifetime of the individuals is only known to fall in an interval, such that $t_i \in (L_i, R_i], i = 1,\ldots,n$, where L_i and R_i are the left and right endpoints of the observed intervals, respectively.

The cure rate models are said to be a parametric or semi-parametric models. In the parametric models, a standard probability distributions such as exponential, weibull, Gompertz and generalized F can be employed. Nonetheless, the main limitation of the parametric cure models is that it is sometimes hard to find a distribution flexible enough to fit the observed data. Therefore, the non-parametric techniques are considered to be more attractive under the violation of the parametric assumptions.

In the following sections, however, the non-parametric techniques are employed to estimate the survival function, based on Turnbull and Kaplan Meier estimators, and followed on to compare the estimation of the cure fraction via a simulation study considering the two non-parametric estimators.

MATERIALS AND METHODS

Turnbull Estimator

In case of right censored data, one can use the Kaplan-Meier estimator to obtain the survival function. However, with interval censored data, this particular estimator is not a suitable one, and it is Turnbull who formulated an algorithm that works on the principle of EM algorithm based on a sample of observed intervals $[L_i, R_i]$, $i = 1, ..., n$, which contains the independent random variables $t_1, ..., t_n$.

For this algorithm, equivalence intervals such as $J_1 = (q_1, p_1]$, $J_2 = (q_2, p_2]$, ..., $J_{\mathrm{m}} = (q_{\mathrm{m}}, p_{\mathrm{m}}]$ must be extracted to determine the jumps $(s_1, s_2, ..., s_m)$ of the cumulative distribution function and hence the survival function. To find the equivalence intervals, consider all the intervals $[L_i, R_i]$ for $i = 1, ..., n$, and order the $2n$ endpoints in ascending order, and each end point "L" that is then immediately followed by the end point "R" which is an equivalence interval.

Let $\alpha_{ij} = 1_{\{[q_j,p_j] \subseteq [L_i,R_i]\}}$, $i = 1, ..., n, j = 1, ..., m$,be the indicator variable of whether or not $[q_j, p_j]$lies within $[L_i, R_i]$. Then, the probability that t_i falls in the $[q_j, p_j]$ given vector of the jumps $s = (s_1, ..., s_m)^T$ is given by:

$$\mu_{ij}(s) = \frac{\alpha_{ij} s_j}{\sum_{j=1}^{m} \alpha_{ij} s_j}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, m$$

Since the survival function is constant outside the intervals $[q_j, p_j]$, the proportion of the observations in $[q_j, p_j]$ is given by:

$$\pi_j(s) = \frac{\sum_{i=1}^{n} \mu_{ij}(s)}{n}, \qquad j = 1, \ldots m$$

The vector $\boldsymbol{s}$ is said to be self consistent if,

$$s_j = \pi_j(s), \; j = 1, \ldots, m$$

Note: to find μ_{ij}, we can use an initialization of **s** such that $s^k = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right), k = 0$, and then follow up to find s_j^{k+1} until stopping conditions such as $\sum_{j=1}^{m} (s_j^{k+1} - s_j^k)^2 < \in, \in > 0$.
Thus, the Turnbull estimator of the survival function can be defined as follows:

$$\hat{S}(t) = \begin{cases} 1 & : \; if \; t_i < q_1 \\ 1 - \sum_{k=1}^{j} s_k & : \; if \; p_j \leq t_i < q_{j+1}, \qquad j = 1, \ldots, m, \qquad i = 1, \ldots, n \\ 0 & : \; if \; t_i < q_1 \end{cases} \tag{3}$$

See Klein and Moeschberger (2003).

Estimation of the Entire Survival Function

Since the survival function is not observed in the equivalence intervals and hence, the survival function amongst the interval $[L_i, R_i]$ which contains at least one of the equivalent classes is unknown if $t_i \in [q_j, p_j]$. Furthermore, the details about the true lifetime are not available, and the only thing that is known is that it belongs to an observed interval. Then, the lifetime t_i can be generated randomly from the interval $[L_i, R_i]$ when both endpoints are observed, while in the case of right censoring where the right endpoint R_i is not observed, it is possible to substitute this endpoint by the last visit time and to do the generation.

In case the generated life times fall in the equivalence intervals, the survival function can then be defined as follows:

- Generating a sufficient number W of sequential values from the equivalence interval (q_j, p_j), such that $T_{jW} = (T_{j1}, ..., T_{jW})$ and $q_j < T_{j1} < T_{j2}$. $1 \leq j \leq m$.
- Generating W sequential values between the corresponding values of the survival function at the endpoints of the equivalence interval (q_j, p_j), such that $S_{jW} = (S_{j1}, ..., S_{jW})$, and $S_{j1} > S_{j2} ... > S_{jW}$, where $S_{ji} = \lim_{t \to qj-} \hat{S}(t)$, and $S_{jw} = \lim_{t \to pj+} \hat{S}(t)$. Note that $S_{j1} = 1$ if $q_j = 0$, and $S_{jW} = 0$ if $p_j = 0$ for all $j = 1, ..., m$.

Then, the survival function can be defined as follows:

$$\hat{S}_j(t) = \begin{cases} \dfrac{(S_{jk}) + (S_{j(k+1)})}{2} & : \ if\ T_{jk} \leq t_{ji} < T_{j(k+1)} \\ \hat{S}(t) & : \ otherwise \end{cases} \tag{4}$$

$$k = 1, ..., W, \qquad i = 1, ..., n \qquad j = 1, ..., m$$

Kaplan-Meier Estimator

The standard non-parametric estimator of the survival function is the Kaplan-Meier (KM) estimator, which is also known as the product limit estimator. This estimator is defined as follows:

$$\hat{S}(t) = \prod_{tj \leq t} \left(\frac{n_i - d_i}{n_i} \right) = \prod_{i=1}^{n} \left(1 - \frac{d_i}{n_i} \right) \tag{5}$$

where ti $t_i \leq t \leq t_{i+1}$, d_i represents the number of failures at time t such that $J_1 = (q_1, p_1]$, $J_2 = (q_2, p_2]$, ... , $J_m = (q_m, p_m]$, and n_i indicates the number of individuals who have not experienced the interested event, and have also not been censored by time t. From equation (5), it is seen that $\hat{S}(t) = 1$ when t is less than the first failure time, i.e. $t < t_i$.

The Kaplan-Meier estimator estimates the jumps of the survival function at the observed times. The jumps on the survival curve are dependent upon the number of events observed at each event time, and also on the pattern of the censored observations before the event time.

In the case of interval data, using the midpoint of each interval to represent the exact survival time is a common practice amongst the analysts, and then applying the Kaplan-Meier method will yield the estimated survival function. If the right endpoints of some intervals are not specified, i.e. right censored, it is then possible to use the maximum value of the visit times to represent the right endpoint for that interval. However, this procedure may produce invalid inference. Due to the lack of efficient statistical methodology and available software, the Kaplan Meier estimator can be implemented.

Midpoint imputation is only applicable when the time periods between the consecutive visits are short (Law & Brookmeyer, 1992). Thus, when the width of the interval increases, we may run into problems. Furthermore, the standard error of the estimator is underestimated since the midpoint imputation assumes that the failure times are exactly known when in fact they are not, (Kim, 2003).

RESULTS AND DISCUSSION

Let the censoring and cure indicators for interval censored data be as follows:

$$\alpha = \begin{cases} 0: \textit{censored} \\ 1: \textit{otherwise} \end{cases} \qquad c = \begin{cases} 0: \textit{cured} \\ 1: \textit{otherwise} \end{cases}$$

Then, the log likelihood function can be obtained by:

$$L_c = \log \prod\nolimits_{i=1}^{n} [\{(f_i^*)\}(1 - e^{-\theta})^{c_i}]^{\alpha_i} [\{e^{-\theta}\}^{1-c_i} \{(1 - e^{-\theta}) S_i^*\}^{c_i}]^{1-\alpha_i} \tag{6}$$

where S_i^* is the survival function of the censored-uncured patients which might be evaluated using the Turnbull or Kaplan Meier estimators, and f_i^* is the probability density function of the uncensored individuals.

One of the most attractive features of the BCH model is that it can be written as a mixture model, where the survival function can be obtained using:

$$\begin{aligned} S(t) &= \exp(-\theta F(t)) \\ &= \exp(-\theta) + [1 - \exp(-\theta)] \left[\frac{\exp(-\theta F(t)) - \exp(-\theta)}{1 - \exp(-\theta)} \right] \end{aligned}$$

Comparing this formula with the mixture model in Equation (1), the survival function of the uncured patients S_i^* can then be represented by:

$$\begin{aligned} S_i^* &= \frac{\exp(-\theta F(t_i)) - \exp(-\theta)}{1 - \exp(-\theta)} \\ &= \frac{\exp(\theta[1 - F(t_i)]) - 1}{\exp(\theta) - 1} \end{aligned}$$

since $S(t_i) = 1 - F(t_i)$, then

$$S_i^* = \frac{\exp(\theta S(t_i)) - 1}{\exp(\theta) - 1} \tag{7}$$

where $S(t_i)$ is the survival function for the i^{th} censored individual.

Furthermore, the probability density function f_i^* can be estimated using the jumps of the survival function which can be obtained by:

$$f_i^* = M_i = F(R_i) - F(L_i), \qquad i = 1, 2, \ldots, n$$

where $F(R_i)$ and $F(L_i)$ are the cumulative distribution functions at the endpoints of the observed interval. Therefore, the log likelihood function can be re-written as follows:

$$\begin{aligned} L_c &= \log \prod\nolimits_{i=1}^{n} [\{(M_i)(1 - e^{-\theta})\}^{c_i}]^{\alpha_i} \cdot \left[\{e^{-\theta}\}^{1-c_i} \left\{ (1 - e^{-\theta}) \left(\frac{e^{\theta S(t_i)} - 1}{e^{\theta} - 1} \right) \right\}^{c_i} \right]^{1-\alpha_i} \\ &= \sum\nolimits_{i=1}^{n} \alpha_i c_i \log(M_i) - \theta \sum\nolimits_{i=1}^{n} (1 - \alpha_i)(1 - c_i) + \log(1 - e^{-\theta}) \sum\nolimits_{i=1}^{n} c_i \\ &\quad + \sum\nolimits_{i=1}^{n} c_i (1 - \alpha_i) \log \left[\frac{e^{\theta S(t_i)} - 1}{e^{\theta} - 1} \right] \end{aligned} \tag{8}$$

Maximizing L_c is subjected to the condition $\sum_{i=1}^{n} M_i \leq 1$. Let q be a non-negative slack variable i.e. $\sum_{i=1}^{n} M + q = 1$. By adding the Lagrange multiplier λ, the log likelihood function can then be re-written as follows:

$$\begin{aligned} L_c = {} & \sum_{i=1}^{n} \alpha_i c_i \log(M_i) - \theta \sum_{i=1}^{n} (1-\alpha_i)(1-c_i) + \log(1-e^{-\theta}) \sum_{i=1}^{n} c_i \\ & + \sum_{i=1}^{n} c_i (1-\alpha_i) \log\left[\frac{e^{\theta S(t_i)} - 1}{e^{\theta} - 1}\right] - \lambda[(\sum_{i=1}^{n} M_i + q) - 1] \end{aligned} \tag{9}$$

The solution of the following equations is the desired estimates of the parameters:

$$\frac{\partial L_c}{\partial \theta} = 0, \frac{\partial L_c}{\partial M_i} = 0, \frac{\partial L_c}{\partial \lambda} = 0, \frac{\partial L_c}{\partial q} = 0$$

such that $\frac{\partial L_c}{\partial \theta} = 0$ implies

$$\begin{aligned} \frac{\partial L_c}{\partial \theta} = {} & \left(\frac{e^{-\theta}}{1-e^{-\theta}}\right) \sum_{i=1}^{n} c_i - \sum_{i=1}^{n} (1-\alpha_i)(1-c_i) \\ & + \sum_{i=1}^{n} c_i (1-\alpha_i) \left[\frac{(e^{\theta}-1)(S(t_i) e^{\theta S(t_i)}) - e^{\theta}(e^{\theta S(t_i)} - 1)}{(e^{\theta S(t_i)} - 1)(e^{\theta} - 1)}\right] \end{aligned} \tag{10}$$

which can be simplified as follows:

$$\begin{aligned} & \sum_{i=1}^{n} c_i - e^{\theta} \sum_{i=1}^{n} (1-\alpha_i) + \sum_{i=1}^{n} (1-\alpha_i)(1-c_i) \\ & + (e^{\theta} - 1) \sum_{i=1}^{n} c_i (1-\alpha_i) \left[\frac{(S(t_i))}{(1 - e^{-\theta S(t_i)})}\right] = 0 \end{aligned} \tag{11}$$

Similarly, $\frac{\partial L_c}{\partial M_i} = 0, i = 1, \ldots, n$ implies:

$$\sum_{i=1}^{n} \alpha_i c_i \frac{1}{M_i} - n\lambda = 0 \tag{12}$$

$\frac{\partial L_c}{\partial \lambda} = 0$ also implies:

$$\sum_{i=1}^{n} M_i + q - 1 = 0 \tag{13}$$

$\frac{\partial L_c}{\partial q} = 0$ implies:

$$\lambda = 0 \tag{14}$$

The solution of equation (11) is our desired estimate of θ, but c_i is partially missing and so the EM algorithm is necessary.

The EM Algorithm

Suppose that the data set is given in the form ($[L_i, R_i]$, α_i), $i = 1, 2, ..., n$, where $[L_i, R_i]$ denotes the observed interval that includes the i^{th} patient lifetime, and α_i is the censoring indicator. The cure indicator c_i is partially missing and this will be handled in the EM algorithm.

However, for the *m* uncensored individuals α_i and c_i, $j = 1, ..., m$, are observed and both are equal to 1, while for $i = m + 1$, α_i is observed and equals to 0 but c_i is not observed and it might be 1 or 0. Thus, in the EM algorithm, the E-step calculates the expectation of (8) given the observed data set. The expected value of the log likelihood function can be represented by:

$$E[L_c] = E_1[L_c/a_j, c_j] + E_2[L_c/\alpha_i]$$

The expected value of the log likelihood function basically depends on $E_2[L_c/\alpha_i]$ which can be defined as follows:

$$\begin{aligned} E_2[L_c/\alpha_i] &= -\theta\sum_{i=m+1}^{n}(1-c_i) + \log(1-e^{-\theta})\sum_{i=m+1}^{n}c_i + \sum_{i=m+1}^{n}c_i\log\left(\frac{e^{\theta S(t_i)}-1}{e^{\theta}-1}\right) \\ &= -\theta\sum_{i=m+1}^{n}(1-c_i) + \sum_{i=m+1}^{n}c_i\left[\log\left(\frac{e^{\theta S(t_i)}-1}{e^{\theta}-1}\right)\right] \end{aligned} \tag{15}$$

Where $\sum_{i=m+1}^{n}(1-c_i)$, $\sum_{i=m+1}^{n}c_i\log\left(\frac{e^{\theta S(t_i)}-1}{e^{\theta}-1}\right)$ and $\sum_{i=m+1}^{n}c_i$ are the sufficient statistics.

Peng and Carriere (2002) defined g_i as the expected value of the i^{th} patient to be uncured conditional on the current estimates of α_i and the survival function of uncured patients S_i^*, such that:

$$g_i = \alpha_i + (1-\alpha_i)\left[\frac{[1-e^{-\theta}]S_i^*}{[e^{-\theta}]+[1-e^{-\theta}]S_i^*}\right]$$

For simplicity, let $p_i = E(1 - c_i) = 1 - g$, $i = m + 1$, which indicates the expected value of the i^{th} patient to be cured such that for censored individuals:

$$p_i = 1 - g_i = 1 - \left[\frac{[1-e^{-\theta}]S_i^*}{[e^{-\theta}]+[1-e^{-\theta}]S_i^*}\right] = 1 - \left[\frac{[1-e^{-\theta}]\left(\frac{e^{\theta S(t_i)}-1}{e^{\theta}-1}\right)}{[e^{-\theta}]+[1-e^{-\theta}]\left(\frac{e^{\theta S(t_i)}-1}{e^{\theta}-1}\right)}\right] = -e^{\theta S(t_i)} \tag{16}$$

Using these notations, the sufficient statistics can then be re-written as follows:

$$\begin{aligned} e_1 &= \sum_{i=m+1}^{n}E(1-c_i) = \sum_{i=m+1}^{n}p_i = \sum_{i=m+1}^{n}e^{-\theta S(t_i)} \\ e_2 &= \sum_{i=m+1}^{n}E(c_i) = \sum_{i=m+1}^{n}(1-p_i) = \sum_{i=m+1}^{n}(1-e^{-\theta S(t_i)}) \\ e_3 &= \sum_{i=m+1}^{n}E\left[c_i\log\left(\frac{e^{\theta S(t_i)}-1}{e^{\theta}-1}\right)\right] = \sum_{i=m+1}^{n}(1-p_i)\log\left(\frac{e^{\theta S(t_i)}-1}{e^{\theta}-1}\right) \end{aligned}$$

In the light of this equation (11), it can be re-written as follows:

$$\begin{aligned} &\left[\sum_{j=1}^{m}c_j + \sum_{i=m+1}^{n}c_i\right] - e^{\theta}\left[\sum_{j=1}^{m}(1-\alpha_j) + \sum_{i=m+1}^{n}(1-\alpha_i)\right] + \\ &\left[\sum_{j=1}^{m}(1-\alpha_j)(1-c_j)\right] + \sum_{i=m+1}^{n}(1-\alpha_i)(1-c_i) + \\ &(e^{\theta}-1)\left[\sum_{j=1}^{m}c_j(1-\alpha_j)\left(\frac{S(t_j)}{1-e^{-\theta S(t_j)}}\right) + \sum_{i=m+1}^{n}c_i(1-\alpha_i)\left(\frac{S(t_j)}{1-e^{-\theta S(t_j)}}\right)\right] = 0 \end{aligned} \tag{17}$$

However, for some initial values of (θ^t) solve for e_1, e_2 and p_i then θ^{t+1} is the numerical solution of equation (17) with respect to θ. Repeat until stopping condition such as $\theta^{t+1} + \theta^t$, ε is small positive value (e.g. 0.0001).

Simulation Study

In simulation studies based on survival analysis, many common distributions can be used to generate the failure time data sets, where the most common distributions that might be employed in such studies are the exponential and Weibull distribution since they fit the data very well. However, in this simulation and to control the data generation process, the exponential distribution with various values of the scale parameter λ has been considered, where λ can be replaced by the values 0.5, 1, 1.5 and 2 respectively which imply various censoring rates for the generated data sets. For each assigned value of λ, a 100 data sets were generated such that each data set comprised 100 observations. The steps used for data generation are as follows (Goulin *et al.,* 2008):

(a) Generate the true survival time t from an exponential distribution using the proposed values of the scale parameter.

(b) Generate a vector V for the clinic visits, assuming that there are 20 clinic visits, in case of exponential distribution, the first visit v_1 was generated from $U(0,0.115)$, and then the next visit v_2 was generated from $U(v_1,v_1 + 0.115)$. The other visit times were generated in the same manner. A uniform distribution is considered in such case to regulate the times of the clinic visits and hence gain short and equivalent lengths of the intervals.

(c) Generate a 100 × 2 empty matrix named "bound" for each data set. The entries of bound matrix are the intervals endpoints for each individual after comparing the true survival time with the 20 visit times. In case of right censoring the right end point is replaced by "*Inf*". The formula used for end points determination is:

For $i = 1, ..., 100, j = 1, .., 20$

$$bound[i,1] = \begin{cases} 0 : if \ \ t[i] < V[1] \\ V[j] : if \ \ V[j] < t[i] < V[j+1] \\ V[20] : if \ \ t[i] > V[20] \end{cases}$$

$$bound[i,2] = \begin{cases} V[1] : if \ \ t[i] < V[1] \\ V[j+1] : if \ \ V[j] < t[i] < V[j+1] \\ Inf \ : t[i] > V[20] \end{cases}$$

(d) Generate a 100 × 2 empty matrix named "status" based on the "bound" matrix and the "status" matrix can then be defined as follows:

$$\text{status } [i,1] \equiv \text{censoring indicator } \alpha_i = \begin{cases} 0 \ : \ if \ bound\,[i,2] = Inf \\ 1 \ : \ if \ \ otherwise \end{cases}$$

status $[i,1] \equiv$ cured indicator $c_i = \begin{cases} 0 & : \ if \ \alpha_i = 0 \\ 1 & : \ if \ \ otherwise \end{cases}$

Note: We assumed that all right censored individuals are cured as a special case.

The Turnbull and Kaplan Meier procedures are employed for each generated data set to estimate the survival function, and hence estimate the cure fraction. In this simulation, the bias of the cure fraction and also the relative efficiency (*RE*) based on the two non-parametric estimators are considered in such that:

$$\text{bias} = \pi - E(\hat{\pi}) \tag{18}$$

and

$$RE = \frac{MSE(Turnbull)}{MSE(KM)} \tag{19}$$

Where, $\hat{\pi}$ is the maximum likelihood estimator for π, and the mean square error $MSE = (biase(\hat{\pi}))^2 + Var(\hat{\pi})$.

A small bias indicates that the estimator is closer to the true value on average and hence more accurate. While *RE* being less than one indicates that the Turnbull estimator is the viable estimator that may be employed to estimate the cure fraction using interval censored data.

Table 1 shows the results of the cure rate estimation based on the two proposed scenarios, where the estimated measures (i.e. Bias, MSE and RE) represent the average of these measures for the whole data sets that have equivalent censoring rate. All the relative efficiency values are less than one, which indicates that the Turnbull estimator in the case of interval censored data and whatever the censoring rate is more efficient than the Kaplan Meier estimator when the midpoint of the observed interval is considered. The bias values obtained from both the

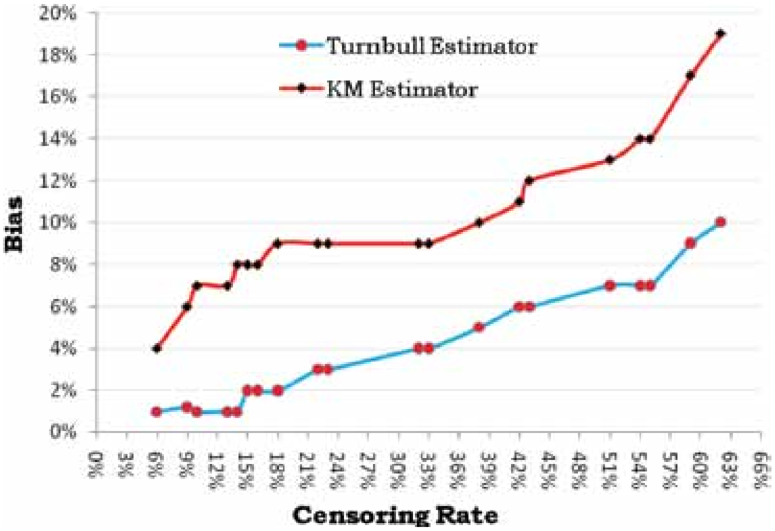

Fig.1: Censoring rates versus bias for Kaplan Meier and Turnbull Estimators

Table 1: Simulation results based on the various values of λ

Run	Censoring Rate	True Cure Rate (**R**)	Turnbull Estimator		Kaplan Meier estimator		Relative Efficiency
			Estimated Cure Rate (**E**)	Bias (**R-E**)	Estimated Cure Rate (**E**)	Bias (**R-E**)	
			λ = 2				
1	6%	6%	5%	1%	2%	4%	0.988
2	9%	9%	8%	1%	3%	6%	0.946
3	10%	10%	9%	1%	3%	7%	0.900
4	13%	13%	12%	1%	6%	7%	0.900
5	14%	14%	13%	1%	6%	8%	0.853
			λ = 1.5				
6	15%	15%	13%	2%	7%	8%	0.864
7	16%	16%	14%	2%	8%	8%	0.864
8	18%	18%	16%	2%	9%	9%	0.816
9	22%	22%	19%	3%	13%	9%	0.832
10	23%	23%	20%	3%	14%	9%	0.832
			λ = 1				
11	32%	32%	28%	4%	23%	9%	0.855
12	33%	33%	29%	4%	24%	9%	0.855
13	38%	38%	33%	5%	28%	10%	0.833
14	42%	42%	36%	6%	31%	11%	0.814
15	43%	43%	37%	6%	31%	12%	0.764
			λ = 0.5				
16	51%	51%	44%	7%	38%	13%	0.748
17	54%	54%	47%	7%	40%	14%	0.700
18	55%	55%	48%	7%	41%	14%	0.700
19	59%	59%	50%	9%	42%	17%	0.636
20	62%	62%	52%	10%	43%	19%	0.591

estimators yield the same indication, and it is noticed that the efficiency of both the estimators declines when the censoring rate goes up, as shown in Fig.1.

CONCLUSION

In this research, two non-parametric estimation methods of the cure fraction were investigated based on the bounded cumulative hazard model using interval censored data. Both the Turnbull and Kaplan Meir estimators were considered, whereby in the case of Kaplan Meir estimator, the midpoint of the intervals could be adopted to represent the exact failure time. The estimation methods were combinations of the straightforward maximum likelihood estimation and the EM algorithm. Hence, the estimating equations were solved numerically since no explicit solutions could be found.

Based on the simulation results and the obtained RE values, however, it was concluded that the Turnbull estimator provides more efficient estimates for the cure fraction using interval censored data compared to the Kaplan Meir estimator. Therefore, based on these results, the analysts who have considered the Kaplan Meier estimator in case of interval censored data should not be too confident with their results. Thus, the Turnbull estimator is recommended to be used for the cure rate estimation rather than the Kaplan Meier estimator.

REFERENCES

Abu Bakar, M. R., Salah, K. A., Ibrahim, N. A., & Haron, K. (2009). Bayesian Approach for Joint Longitudinal and Time-to-Event Data with Survival Fraction. *Bulletin of the Malaysian Mathematical Sciences Society, 32*(1), 75-100.

Aljawadi, B. A., Abu Bakar, M. R., Ibrahim, N. A., & Midi, H. (2011). Parametric Estimation of the Cure Fraction based on BCH Model and Exponential Distribution using Left Censored Data. *Journal of Applied Sciences, 11*(15), 2861-2865.

Andreas, W., Isabella, L., & Antoli, I. (2006). The Modelling of a Cure Fraction in Bivariate Time-to-Event Data. *Australian Journal of Statistics, 35*(1)*,* 67-76.

Berkson, J., & Gage, R. P. (1952). Survival curves for cancer patients following treatment. *Journal of the American Statistical Association, 47*(259)*,* 501-515.

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

Cantor, A. B., & Shuster, J. J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine, 11*, 931-937.

Castro, M. D., Cancho, V. D., & Rodrigues, J. (2010). A hands-on Approach for fitting Long-term Survival Models Under the GSMLSS Framework. *Computer Methods and Programs in Medicine, 97,* 168-177.

Chen, M. H., Ibrahim, J. G., & Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association, 94,* 909-919.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *The Canadian Journal of Statistic, 14*, 257-262.

Gamel, J. W., McLean, I. W., & Rosenberg, S. H. (1990). Proportion cured and mean log survival time as functions of tumour size. *Statistics in Medicine, 9*, 999-1006.

Hanin, L., Tsodikov, A., & Yakovlev, A. (2001). Optimal Schedules of Cancer Surveillance and Tumor Size at Detection. *Mathematical and Computer Modeling 33*, 1419-1430.

Goulin, Z. (2008). *Nonparametric and Parametric survival analysis of censored data with possible violation of method assumptions* (Master thesis dissertaion). University of North Carolina at Greensboro.

Kim, J. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *Journal of the Royal Statistical Society: Series B, 65*(Part 2), 489-502.

Kim, S., Chen, M. H., & Dey, D. K. (2009). A new threshold regression model for survival data with a cure fraction. *Lifetime Data Analysis, 16*, 478-490.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated Data,* (2nd ed.). New York, USA: Springer.

Kuk, A. Y. C., & Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika, 79*, 531-541.

Law, G., Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine, 11*, 1569-1578.

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

Peng, Y., & Carriere, K. C. (2002). An Empirical Comparison of Parametric and Semi-parametric Cure Models. *Biometrical, 44*, 1002-1014.

Peng, Y., & Taylor, J. M. G. (2011). Mixture Cure Model with Random Effects for the Analysis of a Multi-center Tonsil Cancer Study. *Statistics in Medicine, 30*, 211-223.

Rodrigues, J., Cancho, V. G., Castro, M. D., & Louzada-Neto, F. (2009). On the unification of long-term survival models. *Statistics and Probability Letters, 79*, 753-759.

Seppa, K., Hakulinen, T., Kim, H. J., & Laara, E. (2010). Cure Fraction Model with Random Effects for Regional Variation in Cancer Survival. *Statistics in Medicine, 29*, 2781-2793.

Xiang, L., Ma, X., & Yau, K. K. (2011). Mixture Cure Model with Random Effects for Clustered Interval Censored Survival Data. *Statistics in Medicine, 30*, 995-1006.

Yakovlev A. Y., Asselain, B., Bardou, V. J., Fourquet, A., Hoang, T., Rochefediere, A., & Tsodikov, A. D. (1993) . A Simple Stochastic Model of Tumor Recurrence and Its Applications to Data on pre-menopausal Breast Cancer. In B. Asselain, M. Boniface, C. Duby, C. Lopez, J. P.Masson, and J.Tranchefort (Eds.), *Biometrie et Analyse de Dormees Spatio – Temporelles 12* (p. 66-82). ENSA Renned, France: Société Francaise de Biométrie.

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

Zeng, D., Yin, G., & Ibrihim, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association, 101*, 670-684.