Parametric Estimation of the Cure Fraction Based on BCH Model Using Left-Censored Data with Covariates

Bader Ahmad I. Aljawadi (Corresponding author)

Institute for Mathematical Research, Universiti Putra Malaysia

43400 UPM Serdang, Selangor, Malaysia

E-mail: Bader_aljawadi@yahoo.com

Mohd Rizam A. Bakar, Noor Akma Ibrahim & Habshah Midi

Institute for Mathematical Research, Universiti Putra Malaysia

43400 UPM Serdang, Selangor, Malaysia

Abstract

Medical investigations nowadays allow the incorporation of cure individuals in the analysis, especially for chronic diseases such as cancer. Therefore, survival models that incorporate the cured patients in the analysis are called cure rate models. In this paper, we propose an analytical approach for parametric estimation of the cure fraction in cancer clinical trials based on the bounded cumulative hazard (BCH) model with covariates involved in the data set. The analysis is constructed by means of the exponential distribution in the case of left censoring and within the framework of the expectation maximization (EM) algorithm. The analysis provided the analytical solution and a simulation study for the cure rate parameter.

Keywords: Cure fraction, BCH model, Covariates, Left censoring, EM algorithm

1. Introduction

In cancer clinical trials, the population of patients is considered as a heterogeneous population since it is eventually divided into two categories. One group consists of patients who will never experience the event of concern and who are hence considered as cured while the other group comprises patients who remain uncured. However, the main interest of cancer trials is estimating the proportion of cured patients which stands an important criterion for elucidating the trends in the survival of cancer patients. Therefore, the survival models which incorporate the cure fractions in the analysis are called cure rate models.

The first created cure model, which is still widely used in survival analysis, is the model constructed by Boag in 1949 and later developed by Berkson and Gage in 1952. This model is called the mixture cure model since it can estimate the proportion of patients cured as well as the survival function of the uncured patients. According to this model, the survival distribution function can be written in terms of the 'mixture' of the cured, plus the uncured, patients such that:

$$S(t) = \pi + (1 - \pi)S_u(t) \quad . \tag{1}$$

This model is described as parametric or semi-parametric model, depending on whether standard probability distributions are, or are not, employed. If a standard probability distribution like the exponential, Weibull, Gompertz negative binomial, or the generalized F distribution is used, then the model is parametric. If, on the other hand, the mixture model is used without any standard probability distribution, then the model is described as semi-parametric model.

An elaborate literature on the mixture, also known as the standard cure rate model, is found in the works of Goldman (1984), Farewell (1986), Gamel *et al.* (1990), Kuk and Chen (1992), Taylor (1995), Peng and Dear (2000), Sy and Taylor (2000), Peng and Carriere (2003), Mohamad Tajuddin *et al.* (2006), and Abu Bakar *et al.* (2009), amongst others.

In model (1), $S(t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution function. Furthermore, $F(0) = 0$ and $F(\infty) = 1$, so that $S(0) = 1$ and $S(\infty) = \pi$, corresponding to the plateau value. Moreover, the hazard function concomitant to this model is

$$h(t) = \frac{f(t)}{S(t)} \ ,$$

where $f(t)$ is the density function attendant to $F(t)$.

Hence, even mixture cure models fit cancer data well and they usually cannot be viewed literally as describing a mixture of both cured and uncured individuals. However, it should be highlighted that the literal interpretation of the cure model is meaningful in some non-cancer applications, e.g., Hauek *et al.* (1997).

Despite the vast volume of literature on the mixture cure rate model, it has some drawbacks as was illustrated by Chen (Chen *et al.*, 1999). The main drawback is that when a set of covariates is included through, then this model will lack a proportional hazard structure. Therefore, Chen proposed an alternative model which overcomes the drawbacks of the mixture cure model. This alternative model is the "Bounded Cumulative Hazard (BCH)" model which was initially developed by Yakovlev and co-wrkers in 1993 (Yakovlev *et al.*, 1993).

2. Materials and Methods

2.1 The BCH model

The milestone of the (BCH) model is the assumption of an individual in the population left with N cancer cells after the initial treatment. The cancer cells (often called clonogens) grow rapidly and produce a detectable cancer mass later on. The variable N is not observed and has the Poisson, Bernoulli or negative binomial distribution (Rodrigues *et al.*, 2009). Chen *et al.* (1999) considered the Poisson distribution for N with a mean of θ, and in this paper we adopt Chen's assumption.

Given $N = n$, let $Z_i, i = 1,2, \dots n$ be independent random variables with a common distribution function $F(t) = 1 - S(t)$ that is independent of N. The variable Z_i denotes the time which the ith clonogen takes to produce a detectable cancer mass. Then, the time it takes cancer to relapse can be defined by the random variable T where $T = \min\{Z_i, 0 \le i \le N\}$, such that $P(Z_0 = \infty) = 1$, the $Z_i's$ are independent and identically-distributed (i.i.d), and N is independent of the sequence $Z_1, Z_2, \dots, Z_N$. Therefore, the survival function for T, and hence for the population, is given by (Aljawadi *et al.*, 2011):

$$
\begin{aligned}
S(t) &= P(T > t) && \text{(Probability of no cancer by the time t).} \\
&= P(N = 0) + P(Z_1 > t, Z_2 > t \dots, Z_N > t, N \ge 1) \\
&= \exp(-\theta)\left[\sum_{n=0}^{N} \frac{(S(t)\theta)^n}{n!}\right] \\
&= \exp(-\theta)\exp(\theta S(t)) \\
&= \exp(-\theta F(t) \ .
\end{aligned}
\tag{2}
$$

Since $S(\infty) = \exp(-\theta)$ and $F(\infty) = 1$, then model (2) is an improper survival function. The study defines the cure fraction (π) as follows:

$$\pi = S(\infty) = P(N = 0 = \exp(-\theta)) \ . \tag{3}$$

As $\theta \to \infty$, $\pi \to 0$, whereas as $\theta \to 0$, $\pi \to 1$,   *i.e.,* $0 \le \pi \le 1$.

It should be underlined that the first partial derivative of $S(t)$ with respect to t is

$$\frac{\partial S}{\partial t} = -\theta f(t)\exp(-\theta F(t)) \ .$$

Since $1 - S(t) = F(t)$, and accordingly $-\frac{\partial S}{\partial t} = f(t)$, then $\frac{\partial S}{\partial t}$ is an improper survival function , and therefore

$f(t)$ is an improper probability distribution function (p.d.f), as well.

2.2 The likelihood Function

In this analysis, the likelihood function is considered using the left censoring type of input data. In order to analyze such data, let α_i be the indicator of left censoring and c_i be a cure indicator where for the ith patient

$$
\alpha_i = \begin{cases} 0 : censored \\ \\ 1 : otherwise \end{cases}
\quad \text{and} \quad
c_i = \begin{cases} 0 : cured \\ \\ 1 : otherwise \end{cases} \ .
$$

If $\alpha_i = 1$, then $c_i = 1$, but if $\alpha_i = 0$, then c_i is not observed and it can be either one or zero, assuming that censoring is independent of failure times.

Suppose that T is a random variable with probability density function $f(t; \theta)$, θ to be estimated and $t_1, t_2, \ldots, t_n$ is a random sample of size n, then the joint probability density function is given as ;

$$L(t_1, t_2, \ldots, t_n; \theta) = \prod_{i=1}^{n} f(t_i; \theta). \tag{4}$$

In parametric maximum likelihood estimation of the cumulative distribution function, both $F(.)$ and the probability density function $f(.)$ for the entire population are known. This study employs the exponential distribution for $S_u(t_i)$ and $f_u(t_i)$ such that:

$$S_u(t) = e^{-\lambda t} \quad \text{and} \quad f_u(t) = \lambda e^{-\lambda t}.$$

However, in the case of left censoring the survival function of the uncured patients becomes

$$S_u(t) = 1 - e^{-\lambda t}.$$

Given α_i and c_i (i.e. the complete data are available), then the complete log likelihood function is

$$l_c = \log \prod_{i=1}^{n} [\{f_u(t_i)(1-\pi)\}^{c_i}]^{\alpha_i} [\{\pi\}^{1-c_i}\{(1-\pi)(S_u(t_i))\}^{c_i}]^{1-\alpha_i}$$

where $f_u(t_i)$ and $S_u(t_i)$ are the p.d.f and the survival function for the uncured patients, respectively. Consequently, the log-likelihood function becomes:

$$l_c = \log \prod_{i=1}^{n} [\{(\lambda e^{-\lambda t_i})(1-e^{-\theta})\}^{c_i}]^{\alpha_i} [\{e^{-\theta}\}^{1-c_i}\{(1-e^{-\theta})(1-e^{-\lambda t_i})\}^{c_i}]^{1-\alpha_i}. \tag{5}$$

When covariates are involved in the analysis, the scale parameter of the exponential distribution (λ) given the jth covariates can be expressed as:

$\lambda = \exp(B^T V_{ij})$, where V_{ij} and B^T are the covariates and coefficients vectors, respectively; $i = 1, \ldots, n$; and $j = 1, \ldots, J$.

Therefore, the log-likelihood function defined in equation (5) becomes:

$$l_c = \log \prod_{i=1}^{n} \left[\left\{ (\exp(B^T V_{ij}) e^{-\exp(B^T V_{ij})t_i})(1-e^{-\theta}) \right\}^{c_i} \right]^{\alpha_i} \times$$

$$\left[\{e^{-\theta}\}^{1-c_i} \left\{ (1-e^{-\theta})(1-e^{-\exp(B^T V_{ij})t_i}) \right\}^{c_i} \right]^{1-\alpha_i}$$

$$= \sum_{i=1}^{n} \alpha_i c_i [B^T V_{ij}] - \sum_{i=1}^{n} \alpha_i c_i [t_i \exp(B^T V_{ij})] + \log(1-e^{-\theta}) \sum_{i=1}^{n} \alpha_i c_i +$$

$$-\theta \sum_{i=1}^{n} (1-\alpha_i)(1-c_i) + \log(1-e^{-\theta}) \sum_{i=1}^{n} c_i(1-\alpha_i) +$$

$$\sum_{i=1}^{n} c_i(1-\alpha_i) \log[1-e^{-\exp(B^T V_{ij})t_i}]$$

$$= \sum_{i=1}^{n} \alpha_i c_i B^T V_{ij}[1-t_i] - \theta \sum_{i=1}^{n} (1-\alpha_i)(1-c_i) + \log(1-e^{-\theta}) \sum_{i=1}^{n} c_i +$$

$$\sum_{i=1}^{n} c_i(1-\alpha_i) \log[1-e^{-\exp(B^T V_{ij})t_i}]. \tag{6}$$

The solutions of $\frac{\partial l_c}{\partial \theta} = 0$ and $\frac{\partial l_c}{\partial B_1} = 0, \ldots, \frac{\partial l_c}{\partial B_J} = 0$ are our desired estimates of θ and B^T, where

$$\frac{\partial l_c}{\partial \theta} = -\sum_{i=1}^{n} (1-\alpha_i)(1-c_i) + \frac{(e^{-\theta})(\sum_{i=1}^{n} c_i)}{1-e^{-\theta}} = 0$$

$$= -\sum_{i=1}^{n} (1-\alpha_i)(1-c_i) + \frac{(\sum_{i=1}^{n} c_i)}{e^{\theta}-1} = 0$$

$$\Rightarrow \theta = \log\left[\frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} (1-\alpha_i)(1-c_i)} + 1\right]. \tag{7}$$

$$\frac{\partial l_c}{\partial B_1} = \sum_{i=1}^{n} V_{1i} \alpha_i c_i[1-t_i] + \sum_{i=1}^{n} c_i(1-\alpha_i)\left[\frac{\left(-e^{-\exp(B^T V_{ij})t_i}\right)\left(-e^{(B^T V_{ij})t_i}\right)(t_i V_{i1})}{1-e^{-\exp(B^T V_{ij})t_i}}\right] = 0$$

$$\vdots$$

$$\frac{\partial l_c}{\partial B_J} = \sum_{i=1}^{n} V_{Ji} \alpha_i c_i[1-t_i] + \sum_{i=1}^{n} c_i(1-\alpha_i)\left[\frac{\left(-e^{-\exp(B^T V_{ij})t_i}\right)\left(-e^{(B^T V_{ij})t_i}\right)(t_i V_{ij})}{1-e^{-\exp(B^T V_{ij})t_i}}\right] = 0$$

$$\tag{8}$$

As the cure status c_i is not fully observed, the EM algorithm will be used. Before employment of the EM algorithm, g_i is defined as the expected value of the ith patient to be uncured under the conditions of current

estimates of α_i and the survival function of uncured patients, $S_u(t_i)$, and its value is drawn from the equation (Peng and Dear, 2000):

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

For censored individuals $\alpha_i = 0$ and hence the equation giving g_i can be re-written as follows:

$$g_i = \left[\frac{[1 - e^{-\theta}]S_u(t_i)}{[e^{-\theta}] + [1 - e^{-\theta}]S_u(t_i)}\right]$$

$$= \left[\frac{[1 - e^{-\theta}]\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]}{[e^{-\theta}] + [1 - e^{-\theta}]\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]}\right] \ . \tag{9}$$

For simplicity, we can define p_i as the probability of cured individuals such that

$$p_i = E(1 - c_i) = 1 - g_i \quad ,$$

$$= 1 - \left[\frac{[1 - e^{-\theta}]\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]}{[e^{-\theta}] + [1 - e^{-\theta}]\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]}\right]$$

$$= \left[\frac{1}{1 + [e^{\theta} - 1]\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]}\right] \ . \tag{10}$$

2.3 The EM algorithm

Suppose that the data vector is in the form of $(t_i, \alpha_i, c_i, V_j)$. For $i = 1, \ldots, n$ and $j = 1, \ldots, J$, If m represents the number of uncensored individuals, then $(n - m)$ is the number of censored individuals. Accordingly, the observed data are (i) the lifetime (t_i); (ii) the censoring status ($\alpha_i = 1$) for $i = 1 \ldots n$; (iii) the cure status ($c_i = 1$) for $i = 1 \ldots m$, $\forall\, m < n$, and (iv) the covariates vector (V_j) for all j, while the only unobserved data are the cure status c_i for $i = (m + 1) \ldots n$.

However, in the expectation step (E-Step) we determine the expected value of the log likelihood function (6) as follows:

$$E(l_c/\alpha_i, c_i, t_i, V_i) = \sum_{i=1}^{m} \alpha_i c_i B^T V_{ij}[1 - t_i] + \sum_{i=m+1}^{n} \alpha_i c_i B^T V_{ij}[1 - t_i] - \theta\sum_{i=1}^{m}(1 - \alpha_i)(1 - c_i) -$$

$$\theta\sum_{i=m+1}^{n}(1 - \alpha_i)(1 - c_i) + \log\left(1 - e^{-\theta}\right)\sum_{i=1}^{m} c_i + \log\left(1 - e^{-\theta}\right)\sum_{i=m+1}^{n} c_i +$$

$$\sum_{i=1}^{m} c_i(1 - \alpha_i)\log\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right] + \sum_{i=m+1}^{n} c_i(1 - \alpha_i)\log\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]$$

$$= \sum_{i=1}^{m} B^T V_{ij}[1 - t_i] - \theta\sum_{i=m+1}^{n}(1 - c_i) + m\log\left(1 - e^{-\theta}\right) +$$

$$\log\left(1 - e^{-\theta}\right)\sum_{i=m+1}^{n} c_i + \sum_{i=m+1}^{n} c_i\log\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]$$

$\sum_{i=m+1}^{n}(1 - c_i)$, $\sum_{i=m+1}^{n} c_i$ and $\sum_{i=m+1}^{n} c_i\log\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]$ are the sufficient statistics for the parameters' vector $(B^T, \theta)^T$. It follows that the log-likelihood based on complete data is linear in the complete data sufficient statistics, and then the E-step requires the computation of $E_{B^T,\theta}(\sum_{i=m+1}^{n}(1 - c_i))$, $E_{B^T,\theta}(\sum_{i=m+1}^{n} c_i)$ and $E_{B^T,\theta}\left(\sum_{i=m+1}^{n} c_i\log\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]\right)$.

Based on the probability of cure of patients given by equation (10) let:

$$e_1 = E_{B^T,\theta}\left(\sum_{i=m+1}^{n}(1 - c_i)\right) = \sum_{i=m+1}^{n}(p_i)$$

$$= \sum_{i=m+1}^{n}\left[\frac{1}{1 + [e^{\theta} - 1]\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]}\right] \tag{11}$$

$$e_2 = E_{B^T, \theta}\left(\sum_{i=m+1}^{n} c_i\right) = \sum_{i=m+1}^{n}(1-p_i)$$

$$= \sum_{i=m+1}^{n}\left[1 - \frac{1}{1+[e^{\theta}-1]\left[1-e^{-\exp\left(B^T V_{ij}\right)t_i}\right]}\right] \tag{12}$$

and

$$e_3 = E_{B^T, \theta}\left(\sum_{i=m+1}^{n} c_i \log\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right]\right)$$

$$= \sum_{i=m+1}^{n}\left[\frac{[1-e^{-\theta}]\left[1-e^{-\exp\left(B^T V_{ij}\right)t_i}\right]}{[e^{-\theta}]+[1-e^{-\theta}]\left[1-e^{-\exp\left(B^T V_{ij}\right)t_i}\right]}\right]\log\left[1 - e^{-\exp\left(B^T V_{ij}\right)t_i}\right] \tag{13}$$

On grounds of the data partition defined above and the sufficient statistics, we can re-write equation (7) and the system of nonlinear equations in (8) respectively as follows:

$$\theta = \log\left[\frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n}(1-\alpha_i)(1-c_i)} + 1\right] = \log\left[\frac{\sum_{i=1}^{m} c_i + \sum_{i=m+1}^{n} c_i}{\sum_{i=1}^{m}(1-\alpha_i)(1-c_i)+\sum_{i=m+1}^{n}(1-\alpha_i)(1-c_i)} + 1\right]$$

$$= \log\left[\frac{m+\sum_{i=m+1}^{n} c_i}{\sum_{i=m+1}^{n}(1-c_i)} + 1\right]$$

$$= \log\left[\frac{m+e_2}{e_1} + 1\right]. \tag{14}$$

$$\frac{\partial l_c}{\partial B_1} = \sum_{i=1}^{m} V_{1i}\alpha_i c_i[1-t_i] + \sum_{i=m+1}^{n} V_{1i}\alpha_i c_i[1-t_i] +$$

$$\sum_{i=1}^{m} c_i(1-\alpha_i)\left[\frac{\left(-e^{-\exp\left(B^T V_{ij}\right)t_i}\right)\left(-e^{(B^T V_{ij})t_i}\right)(t_i V_{i1})}{1-e^{-\exp\left(B^T V_{ij}\right)t_i}}\right] +$$

$$\sum_{i=m+1}^{n} c_i(1-\alpha_i)\left[\frac{\left(-e^{-\exp\left(B^T V_{ij}\right)t_i}\right)\left(-e^{(B^T V_{ij})t_i}\right)(t_i V_{i1})}{1-e^{-\exp\left(B^T V_{ij}\right)t_i}}\right] = 0$$

$$\vdots$$

$$\frac{\partial l_c}{\partial B_J} = \sum_{i=1}^{m} V_{Ji}\alpha_i c_i[1-t_i] + \sum_{i=m+1}^{n} V_{Ji}\alpha_i c_i[1-t_i] +$$

$$\sum_{i=1}^{m} c_i(1-\alpha_i)\left[\frac{\left(-e^{-\exp\left(B^T V_{ij}\right)t_i}\right)\left(-e^{(B^T V_{ij})t_i}\right)(t_i V_{i1})}{1-e^{-\exp\left(B^T V_{ij}\right)t_i}}\right] +$$

$$\sum_{i=m+1}^{n} c_i(1-\alpha_i)\left[\frac{\left(-e^{-\exp\left(B^T V_{ij}\right)t_i}\right)\left(-e^{(B^T V_{ij})t_i}\right)(t_i V_{i1})}{1-e^{-\exp\left(B^T V_{ij}\right)t_i}}\right] = 0$$

$$\frac{\partial l_c}{\partial B_1} = \sum_{i=1}^{m} V_{1i}[1-t_i] + \sum_{i=m+1}^{n}(1-p_i)\left[\frac{\left(-e^{-\exp\left(B^T V_{ij}\right)t_i}\right)\left(-e^{(B^T V_{ij})t_i}\right)(t_i V_{i1})}{1-e^{-\exp\left(B^T V_{ij}\right)t_i}}\right] = 0$$

$$\vdots$$

$$\frac{\partial l_c}{\partial B_J} = \sum_{i=1}^{m} V_{Ji}[1-t_i] + \sum_{i=m+1}^{n}(1-p_i)\left[\frac{\left(-e^{-\exp\left(B^T V_{ij}\right)t_i}\right)\left(-e^{(B^T V_{ij})t_i}\right)(t_i V_{i1})}{1-e^{-\exp\left(B^T V_{ij}\right)t_i}}\right] = 0 \tag{15}$$

However, for the maximization step (M-Step), the complete data maximum likelihood estimates of the parameters are given by equations (14) and the system of nonlinear equations (15). Thus, for an initial value of θ° the system of non-linear equations (15) with respect to the parameters vector (B^T) using any appropriate numerical method, such as the Newton Raphson method, can be solved. The solution of the non-linear equation in (15) in addition to the initial value of θ° can be used to solve the complete data sufficient statistics given by equations (11), (12) and (13). Substitution of these sufficient statistics in equation (14) will give a new value for θ^t that will be reiterated until convergence and eventually the desired cure fraction is $\pi = \exp(-\theta^{t+1})$.

3. Simulation and results

Simulation studies based on left censored data involve similar steps in comparison with right censoring type. In this study, exponential, binomial and uniform distributions were used to generate the data which is composed of the lifetimes t, censoring, cure status and covariates vector (α, c, Z) respectively.

The steps used to generate a left censored data set are shown below:

a) Generate a variable t for true survival time from an exponential distribution with various scale parameter (λ) values since we were quite interested in varying the censoring rates (P) as this enables identification of the pattern which the cure rate estimation will assume. So, for the sake of flexibility in finding out how the pattern of the cure rate would progress as the censoring rate increased, both slowly and rapidly, we were interested in the scale parameter $\lambda \in [0.5, 2]$.

b) Generate another variable for censoring time from a uniform distribution on the interval $(\min(t), [\min(t) + \max(t)]/2)$ to obtain left censored survival times. This censoring variable is denoted as T.

c) Compare the true survival time t with the censoring time T. Then the lifetimes and censoring indicator α can be defined as follows:

$$time[i] = \begin{cases} T[i] & if \quad t[i] \leq T[i] \\ t[i] & if \quad t[i] > T[i] \end{cases}$$

$$\alpha[i] = \begin{cases} 0 & if \quad time[i] = T[i] \\ 1 & if \quad time[i] = t[i] \end{cases}$$

In this simulation we generated 20 data sets each with 100 individuals where as a special case for each set of cohorts we considered the left censored individuals are cured. Thus, the cure indicator c can be defined in the same manner of censoring indicator where

$$c[i] = \begin{cases} 0 & if \quad time[i] = T[i] \\ 1 & if \quad time[i] = t[i] \end{cases}$$

Regarding the covariates, only two covariates were considered: gender, which derived from a binomial distribution; and type of treatment, i.e., chemotherapy or radiotherapy, which derived from binomial distribution.

In this simulation we were interested in the bias and in the mean square error (MSE), where bias is commonly defined as the difference between the true and the expected values of an estimator as given by:

$$bias = \pi - E(\hat{\pi}) ,$$

where $\hat{\pi}$ is the maximum likelihood estimate for π.

Consequently, the MSE of an estimator is known as the expected squared deviation of the estimated parameter value from the true one. By using a standard notation for a scalar parameter, it can be expressed in the following form:

$$MSE_\pi = variance(\hat{\pi}) + bias^2$$

The simulation was carried out with the built-in random generators in the R statistical software to fulfill the entire simulation and the final results are presented next by Tables 3.1 and Figure 3.1, respectively.

The above results show the results of the parametric estimation of the cure fraction when covariates were included in the analysis. The bias and MSE values for the various given rates of censoring indicate that the proposed method of cure rate estimation was more efficient when the censoring rate was low than when it was high, and that the estimation started to diverge in the case of heavy censoring rates. Hence, increasing the proportion of censored data will distort the estimated parameters, and vice versa.

4. Conclusion

By assuming that both $F_\circ(.)$ and $f_\circ(.)$ are known, the researchers investigated the parametric maximum likelihood estimation equations of the cure fraction. The analysis was conducted by consideration of the left censoring case based on the bounded commutative hazard model with the exponential distribution used to represent the survival function of the uncured patients. Covariates were involved in the analysis via the scale parameter of the exponential function whereby the parametric estimation equations of the parameters can be solved numerically by selection of an appropriate numerical method since the researchers could not find an explicit solution. The results demonstrate that cure fraction estimation based on the proposed procedure was more attractive when censoring rate is low than when it is high.

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. *Bull. Malays. Math. Sci. Soc.,* **32**, 75-100.

Aljawadi, B. A., Bakar, M. R., Ibrahim, N. A. (2011). Non-parametric Estimation of Cure Fraction Using Right Censored Data. *American Journal of Scientific Research,* **14**, 79-87.

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

Chen, M. H., Ibrahim, J. G., and 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. and Rosenberg, S. H. (1990). Proportion cured and mean log survival time as functions of tumour size. *Statistics in Medicine,* **9**, 999-1006.

Goldman, A. I. (1984). Survivorship analysis when cure is a possibility: A Monte Carlo study. *Statistics in Medicine,* **3**, 153-163.

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

Maller, R. and S. Zhou. 1996. *Survival Analysis with long-Term Survivors.* 1st Edition. New York: Wiley.

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

Peng. Y. and Carriere. K., C. (2003). An Empirical Comparison of Parametric and Semiparametric Cure Models. *Biometrical,* **44**. 1002-1014.

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

Sy, J. P. and J. M. Taylor. (2000). Estimation in a Cox proportional hazard cure model. *Biometrics,* **54**, 227-236.

Taj Uddin, M., Islam, M. N. and Ibrahim, Q. I. (2006). An Analytical Approach on Cure Rate Estimation Based on Uncensored Data. *Applied Sciences,* **6(3)**, 548-552.

Taj Uddin, M., Sen. A., Noor, M. S., Islam, M. N. and Chowdhury, Z. I. (2006). An Analytical Approach on Non-parametric Estimation of Cure Rate Based on Uncensored Data. *Journal of Applied Sciences,* **6**, 1258-1264.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biostatistics,* **51**, 237-243.

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

Table 3.1 Results of the simulation based on the scale parameter $\lambda \in [0.5, 2]$ and 20 data sets.

Run	Censoring Rate	Real Cure	Expected Cure($\widehat{\pi}$)	Bias	MSE(×100)
1	18%	18%	17%	1%	0.2194
2	20%	20%	19%	1%	0.2194
3	21%	21%	19%	2%	0.2494
4	23%	23%	20%	3%	0.2994
5	25%	25%	22%	3%	0.2994
6	26%	26%	22%	4%	0.3694
7	27%	27%	22%	5%	0.4594
8	29%	29%	23%	6%	0.5694
9	31%	31%	25%	6%	0.5694
10	32%	32%	25%	7%	0.6994
11	33%	33%	25%	8%	0.8494
12	35%	35%	27%	8%	0.8494
13	36%	36%	27%	9%	1.0194
14	38%	38%	29%	9%	1.0194
15	38%	38%	28%	10%	1.2094
16	40%	40%	28%	12%	1.6494
17	42%	42%	28%	14%	2.1694
18	45%	45%	30%	15%	2.4594
19	47%	47%	32%	15%	2.4594
20	50%	50%	34%	16%	2.7694
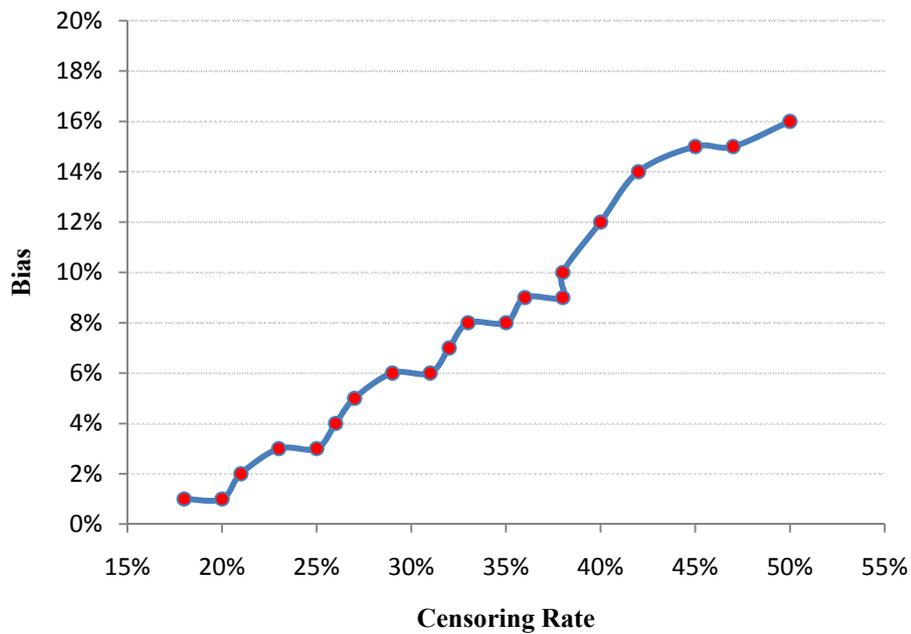
Figure 3.1 Censoring rate versus bias based on the results in table 3.1.