# Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 11 (15): 2861-2865, 2011
ISSN 1812-5654 / DOI: 10.3923/jas.2011.2861.2865
© 2011 Asian Network for Scientific Information

# Parametric Estimation of the Immunes Proportion based on BCH Model and Exponential Distribution using Left Censored Data

[1]Bader Ahmad I. Aljawadi, [2]Mohd Rizam A. Bakar, [2]Noor Akma Ibrahim and [2]Habshah Midi
[1]Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia
[2]Department of Mathematics, Institute for Mathematical Research,
Universiti Putra Malaysia, Malaysia

**Abstract:** In population based cancer clinical trials, a proportion of patients will never experience the interested event and considered as "cured" or "immunes". The majority of recent cancer studies focus on the estimation of immune proportion. In this study we investigated the estimation of proportion of patients curd of cancer in case of left censored data based on the Bounded Cumulative Hazard (BCH) model proposed by Chen in 1999. The analysis provided the Maximum Likelihood Estimation (MLE) of the parameters within the framework of the Expectation Maximization (EM) algorithm where the numerical solutions of the estimation equations of the cure rate parameter could be employed.

**Key words:** Cure fraction, BCH model, left censoring, MLE method, EM algorithm

## INTRODUCTION

Survival models that incorporate the cure fraction in the analysis are called cure rate models. Recently, survival cure models are being widely used in analyzing data from cancer studies. They are used for analyzing survival data from various types of cancer in which a proportion of patients becomes free of any signs or symptoms of the disease, Amiri *et al.* (2008). The first created cure rate model is the mixture model which constructed by Boag (1949) and later developed by Berkson and Gage (1952). In this model, a certain proportion $\pi$ of patients are cured as well as the remain $1-\pi$ are not.

In this model the survival function for the entire population can be written in terms of the 'mixture' of the cured part plus the uncured part such that:

$$S(t) = \pi+(1-\pi) S_u(t) \quad (1)$$

where, $S(t)$ and $S_u(t)$ are the survival functions for the entire population and the uncured patients, respectively. The survival function of uncured patients can be estimated parametrically or non-parametrically which leads to parametric or semi-parametric survival function, respectively, where in the parametric case, a particular distribution for the failure time distribution of uncured patients could be employed such as exponential, Weibull, Gompertz, negative binomial and Generalized F distribution (Savadi-Oskouei *et al.*, 2010).

The literature on mixture cure model could be found in the study of Gamel *et al.* (1990), Kuk and Chen (1992), Taylor (1995), Peng and Dear (2000), Sy and Taylor (2000), Peng and Carriere (2002), Uddin *et al.* (2006), Liu *et al.* (2006a), Yu and Peng (2008) and Abu Bakar *et al.* (2009).

In Eq. 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 $S(\infty) = \pi$ the plateau value. The hazard function concomitant to this model is:

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

where, $f(t)$ is the probability density function (p.d.f) attendant to $F(t)$.

Despite the widely used of the mixture model in survival analysis, it has some limitations as was discussed by Chen *et al.* (1999), some of these drawbacks are:

- The proportional hazard structure which is a desirable property for any survival model cannot be constructed in the presence of covariates
- When including covariates through the parameter $\pi$ via a standard regression model, then mixture model yields improper posterior distributions for many types of non-informative improper priors, including the uniform prior for the regression coefficients
- Mixture model does not appear to describe the underlying biological process generating the failure time, at least the context of cancer relapse, where cure rate models are frequently used

Chen *et al.* (1999) proposed the Bounded Cumulative Hazard (BCH) model developed by Yakovlev *et al.* (1993)

**Corresponding Author:** Bader Ahmad I. Aljawadi, Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia

as the viable alternative to the mixture model. This alternative model is quite attractive for several aspects:

- It is derived from a natural biological motivation
- It has proportional hazard structure through the cure rate parameter
- It is computationally very attractive
- It has a mathematical relationship with the mixture cure rate model

The bounded cumulative hazard model assumes that for an individual in the population left with N cancer cells after the initial treatment. The cancer cells (often called clonogens) grow rapidly and replace the normal tissue later on (cancer relapse). N may follows Poisson, Bernoulli or negative binomial distribution (Rodrigues *et al*., 2009). However, in this study we will consider N to follow the Poisson distribution with a mean of $\theta$.

Let $Z_i$, i = 1, 2, ···N denotes the time of the ith clonogen to produce detectable cancer mass. Then the time it takes cancer to relapse can be defined by the random variable T = min [$Z_i$, $0 \le i \le N$], P ($Z_0 = \infty$) and $Z_i$'s are independent and identically distributed (i.i.d) and that N is independent of the sequence $Z_1, Z_2, \cdots, Z_N$. Therefore, the survival function for T and hence for the population, is given by: S (t) = P (T>t) (Probability no cancer by the time t ).

$$
\begin{aligned}
&= P(N = 0) + P(Z_1 > t, Z_2 > t \cdots, Z_N > t, N \ge 1) \\
&= \exp(-\theta) + [P(Z_1 > t) P(N = 1)] + [P(Z_1 > t) P(Z_2 > t) P(N = 2)] + [P(Z_1 > t) P(Z_2 > t) P(Z_3 > t) P(N = 3)] + \cdots \\
&\quad + [P(Z_1 > t) P(Z_2 > t) \cdots P(Z_n > t) P(N = n)] \\
&= \exp(-\theta) + [S(t) P(N = 1)] + [S(t)^2 P(N = 2)] + [S(t)^3 P(N = 3)] + \cdots [S(t)^n P(N = n)] \\
&= \exp(-\theta) + \sum_{n=1}^{N} [S(t)^n P(N = n)] \\
&= \exp(-\theta) + \sum_{n=1}^{N} \frac{(S(t))^n \exp(-\theta)(\theta)^n)}{n!} \\
&= \exp(-\theta) + \sum_{n=1}^{N} \frac{(S(t)\theta)^n \exp(-\theta)}{n!} \\
&= \exp(-\theta) + \exp(-\theta) \sum_{n=1}^{N} \frac{(S(t)\theta)^n}{n!} \\
&= \exp(-\theta) \left[ 1 + \sum_{n=1}^{N} \frac{(S(t)\theta)^n}{n!} \right] \\
&= \exp(-\theta) \left[ \sum_{n=0}^{N} \frac{(S(t)\theta)^n}{n!} \right] \\
&= \exp(-\theta) \exp(\theta S(t)) \\
&= \exp(-\theta F(t))
\end{aligned} \tag{2}
$$

Since S ($\infty$) = exp (-$\theta$) and F ($\infty$) = 1, then Eq. 2 is an improper survival function. Therefore, the cure fraction $\pi$ can be defined 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 notified that the first derivative of S (t) with respect to t is:

$$\frac{dS}{dt} = \theta f(t) \exp(-\theta F(t))$$

Since 1-S (t) = F (t) and accordingly:

$$-\frac{dS}{dt} = f(t)$$

then ds/dt is an improper survival function and therefore, f (t) is an improper probability density function as well.

## MATERIALS AND METHODS

Suppose that T is a random variable with probability density function f (t; $\theta$), $\theta$ to be estimated and $t_1, t_2, \cdots, t_n$ is a random sample of size n. We are interesting in the likelihood function using the left censored data, because it gives us the possibility to compute the Maximum Likelihood Estimates (MLE) in order to fit a model for censored data. In order to analyze such data let $\alpha_i$ and $c_i$ are indicators for the left censoring and cured, respectively where for the ith patient:

$$\alpha_i = \begin{cases} 0: \text{Censored} \\ 1: \text{Otherwise} \end{cases} \text{and} \; c_i = \begin{cases} 0: \text{Cured} \\ 1: \text{Otherwise} \end{cases}$$

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.

In parametric maximum likelihood method the cumulative distribution function F (.) and the probability density function f (.) for the entire population are known. Thus, given $\alpha_i$ and $c_i$ (i.e., the complete data are available), then the joint probability density function can be written as:

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

Consequently, the complete log likelihood function is:

$$l_c = \log \prod_{i=1}^{n} [\{f_u(t_i)(1-\pi)\}^{c_i}]^{\alpha_i} \left[ \{\pi\}^{1-c_i} \{(1-\pi)(1-S_u(t_i))\}^{c_i} \right]^{1-\alpha_i} \tag{5}$$

where, $f_u(t_i)$ and $S_u(t_i)$ are the p.d.f and the survival function for the uncured patients, respectively.

This study considers the exponential distribution for $S_u(t_i)$ and $f_u(t_i)$ such that:

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

A datum $t_i$ is said to be left-censored if the event occurs at a time before a left bound but it is unknown when it happens, for example, when the date of starting a cancer clinical trial is assigned but for a cancer patient we don't know when the patient has been died. However, in case of left censoring the survival function of the uncured patients becomes $S_u(t) = 1\text{-}e^{-\lambda\tau}$.

Therefore, the log-likelihood function becomes:

$$\begin{aligned}
l_c &= \log\prod_{i=1}^{n}[\{\lambda_e{}^{-\lambda t_i})(1-e^{-\theta})\}^{c_i}]^{\alpha_i}\left[\{e^{-\theta}\}^{1-c_i}\{(1-e^{-\theta})(1-e^{-\lambda t_i})\}^{c_i}\right]^{1-\alpha_i}\\
&= \sum_{i=1}^{n}\log\left[\{(\lambda e^{-\lambda t_i})(1-e^{-\theta})\}^{1-c_i}\right]^{\alpha_i} + \sum_{i=1}^{n}\log\left[\{e^{-\theta}\}^{1-c_i}\{(1-e^{-\theta})(1-e^{-\lambda t_i})\}^{c_i}\right]^{1-\alpha_i}\\
&= \sum_{i=1}^{n}\alpha_i c_i\left[\log\lambda - (\lambda t_i) + \log(1-e^{-\theta})\right] - \theta\sum_{i=1}^{n}(1-\alpha_i)(1-c_i) +\\
&\quad \sum_{i=1}^{n}c_i(1-\alpha_i)[\log(1-e^{-\theta}) + \log(1-e^{-\lambda t_i})]\\
&= \log\lambda\sum_{i=1}^{n}a_i c_i - \lambda\sum_{i=1}^{n}t_i\alpha_i c_i - \theta\sum_{i=1}^{n}(1-\alpha_i)(1-c_i) +\\
&\quad \log[(1-e^{-\theta})\sum_{i=1}^{n}c_i(1-\alpha_i)\log(1-e^{-\lambda t_i})]
\end{aligned} \tag{6}$$

The solutions of:

$$\frac{\partial l_c}{\partial\theta} = 0$$

and:

$$\frac{\partial l_c}{\partial\lambda} = 0$$

are the desired estimates of $\theta$ and $\lambda$, where,

$$\frac{\partial l_c}{\partial\theta} = -\sum_{i=1}^{n}(1-\alpha_i)(1-c_i) + \left(\frac{1}{e^{\theta}-1}\right)\sum_{i=1}^{n}c_i = 0 \tag{7}$$

$$\frac{\partial l_c}{\partial\lambda} = \frac{\Sigma_{i=1}^{n}\alpha_i c_i}{\lambda} - \sum_{i=1}^{n}t_i\alpha_i c_i + \sum_{i=1}^{n}c_i(1-\alpha_i)\left(\frac{t_i}{e^{\lambda t_i}-1}\right) = 0 \tag{8}$$

Solving Eq. 7 implies:

$$\theta = \log\left[\frac{\Sigma_{i=1}^{n}c_i}{\Sigma_{i=1}^{n}(1-\alpha_i)(1-c_i)} + 1\right] \tag{9}$$

While Eq. 8 can be solved numerically since no explicit solution can be found.

As the cure status $c_i$ is not fully observed, the Expectation Maximization (EM) algorithm will employ.

Before implementing the EM algorithm, let's define $g_i$ as the expected value of the ith patient to be uncured conditional on the current estimates of $\alpha_i$ and the survival function of uncured patients, $S_u(t_i)$ (Peng and Dear, 2000):

$$g_i = \alpha_i + (1-\alpha_i)\left[\frac{[1-e^{-\theta}]S_u(t_i)}{[e^{-\theta}]+[1-e^{-\theta}]S_u(t_i)}\right] \tag{10}$$

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

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

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

For simplicity, let $p_i$ to be the probability of cured patients such that $p_i = E(1\text{-}c_t) = 1\text{-}g_i$:

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

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

## THE EM ALGORITHM

EM algorithm is an iterative optimization method which alternates between performing an Expectation (E) step which computes the expectation of the log-likelihood function using the current estimate for the latent variables and Maximization (M) step which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. The EM algorithm (and its faster variant ordered subset expectation maximization) is widely used in many different fields, especially in data clustering in machine learning, computer vision and medicine (Liu *et al*., 2006b; Safarinejadian *et al*., 2009).

However, suppose that the data vector is in the form of $(t_i, \alpha_i, c_i)$. For i = 1···n, the observed data is the lifetime $(t_i)$ and censoring status $(\alpha_i = 1)$ for i = 1···n and also the cure status $(c_i = 1)$, i = 1···m while the unobserved data is the cure status $(c_i)$ for i = (m+1)···n, $\forall m<n$.

In the presence of unobserved data ($c_i$), only a function of the complete-data vector is observed. However, in the E-Step we find the expected value of the log likelihood function given by Eq. 6 as follows:

$$E(l_c / \alpha_i, c_i, t_i) = m\log\lambda - \lambda\sum_{i=1}^{m} t_i - \theta\sum_{i=m+1}^{n}(1-c_i) + m\log(1-e^{-\theta}) + \log(1-e^{-\theta})\sum_{i=m+1}^{n} c_i + \sum_{i=m+1}^{n} c_i \log(1-e^{-\lambda t_i}) \sum_{i=m+1}^{n}(1-c_i), \sum_{i=m+1}^{n} c_i \sum_{i=m+1}^{n}\log(1-e^{-\lambda t_i})$$

are the sufficient statistics for the parameters vector $(\lambda, \theta)^T$.

It follows that the log-likelihood based on complete data is linear in complete data sufficient statistics and then the E-step requires the computation of :

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

and

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

Let:

$$S_1 = E_{\lambda,\theta}\left(\sum_{i=m+1}^{n}(1-c_i)\right) = (n-m)(p_i) = (n-m)\left[\frac{1}{1+[e^{\theta}-1](1-e^{-\lambda t_i})}\right] \quad (11)$$

$$S_2 = E_{\lambda,\theta}\left(\sum_{i=m+1}^{n} c_i\right) = \sum_{i=m+1}^{n}[1-p_i] = \sum_{i=m+1}^{n} c_i\left[1-\frac{1}{1+[e^{\theta}-1](1-e^{-\lambda t_i})}\right] \quad (12)$$

$$S_3 = E_{\lambda,\theta}\left(\sum_{i=m+1}^{n} c_i \log(1-e^{-\lambda t_i})\right) = \sum_{i=m+1}^{n}[1-p_i]t_i = \sum_{i=m+1}^{n} c_i\left[1-\frac{1}{1+[e^{\theta}-1](1-e^{-\lambda t_i})}\right]t_i \quad (13)$$

For the M-Step we can use the complete data maximum likelihood estimates of $(\lambda, \theta)$ given by Eq. 8 and 9 and then substituting the expectations derived in the E-step for the complete data sufficient statistics, such that on grounds of the sufficient statistics, the maximum likelihood equation of $\theta$ implies:

$$\theta^{t+1} = \log\left[\frac{\Sigma_{i=1}^{n} c_i}{\Sigma_{i=1}^{n}(1-\alpha_i)(1-c_i)}\right] = \log\left[\frac{\Sigma_{i=1}^{m} c_i + \Sigma_{i=m+1}^{n} c_i}{\Sigma_{i=1}^{m}(1-\alpha_i)(1-c_i) + \Sigma_{i=1}^{m} c_i(1-\alpha_i)(1-c_i)} + 1\right] = \log\left[\frac{m+S_2}{S_1} + 1\right] \quad (14)$$

While Eq. 8 could be re-rewritten as follows:

$$\frac{\partial l_c}{\partial\lambda} = \frac{m}{\lambda} - \sum_{i=1}^{m} t_i c_i + \sum_{i=m+1}^{n} c_i(1-\alpha_i)\left(\frac{t_i}{e^{\lambda t_i}-1}\right) = 0$$
$$= \frac{m}{\lambda} - \sum_{i=1}^{m} t_i(1-p_i) + \sum_{i=m+1}^{n}(1-p_i)(1-\alpha_i)\left(\frac{t_i}{e^{\lambda t_i}-1}\right) = 0 \quad (15)$$

Thus, the E-step involves evaluating the sufficient statistics given by Eq. 11, 12 and 13 and also $p_i$ using some initial values for the parameters ($\theta^{\circ}$, $\lambda^{\circ}$) followed by M-step involves substituting these values in Eq. 14 and solving Eq. 15 numerically with respect to $\lambda$. The convergence $t^{th}$ iteration is our desired estimates of $\theta$ and $\lambda$ and eventually the desired cure fraction is exp ($-\theta^{t+1}$).

## CONCLUSION

We investigated the maximum likelihood estimation methodology for cure rate estimation based on the bounded cumulative hazard model when the exponential distribution can be used to represent the survival function of the uncured patients. A novel development of the EM algorithm was used to obtain maximum likelihood estimates when the data set has some left censoring observations.

## REFERENCES

Abu Bakar, M.R., K.A. Salah, N.A. Ibrahim and K. Haron, 2009. Bayesian approach for joint longitudinal and time-to-event data with survival fraction. Bull. Malays. Math. Sci. Soc., 32: 75-100.

Amiri, Z., K. Mohammad, M. Mahmoudi, H. Zeraati and A. Fotouhi, 2008. Assessment of gastric cancer survival: Using an artificial hierarchical neural network. Pak. J. Biol. Sci., 11: 1076-1084.

Berkson, J. and R.P. Gage, 1952. Survival curves for cancer patients following treatments. J. Am. Stat. Assoc., 47: 501-515.

Boag, J.W., 1949. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J. R. Stat. Soc., 11: 15-44.

Chen, M.H., J.G. Ibrahim and D. Sinha, 1999. A new Bayesian model for survival data with a surviving fraction. J. Am. Statist. Assoc., 94: 909-919.

Gamel, J.W., I.W. McLean and S.H. Rosenberg, 1990. Proportion cured and mean log survival time as functions of tumour size. Statist. Med., 9: 999-1006.

Kuk, A.Y.C. and C.H. Chen, 1992. A mixture model combining logistic regression with proportional hazard regression. Biometrika, 79: 531-541.

Liu, H., H. Zhong, T. Zhang and Z. Gong, 2006a. A quasi-newton acceleration EM algorithm for OFDM systems channel estimation. Inform. Technol. J., 5: 749-752.

Liu, M., W. Lu and Y. Shao, 2006b. Mixture cure model with an application to interval mapping of quantitative trait loci. LifeTime Data Anal., 12: 421-440.

Peng, Y. and K.B.G. Dear, 2000. A nonparametric mixture model for cure rate estimation. Biometrics, 56: 237-243.

Peng, Y. and K.C. Carriere, 2002. An empirical comparison of parametric and semiparametric cure models. Biometrical. J., 44: 1002-1014.

Rodrigues, J., V.G. Cancho, M. de Castro and F. Louzada-Neto, 2009. On the unification of long-term survival models. Stat. Probab. Lett., 79: 753-759.

Safarinejadian, B., M.B. Menhaj and M. Karrari, 2009. Distributed data clustering using expectation maximization algorithm. J. Applied Sci., 9: 854-864.

Savadi-Oskouei, D., H. Sadeghi-Bazargani, M. Hashemilar and T. DeAngelis, 2010. Symptomatologic versus neuroimaging predictors of in-hospital survival after intracerebral haemorrhage. Pak. J. Biol. Sci., 13: 443-447.

Sy, J.P. and J.M.G. Taylor, 2000. Estimation in a Cox proportional hazard cure model. Biometrics, 56: 227-236.

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

Uddin, M.T., A. Sen, M.S. Noor, M.N. Islam and Z.I. Chowdhury, 2006. An analytical approach on non-parametric estimation of cure rate based on uncensored data. J. Applied Sci., 6: 1258-1264.

Yakovlev, A.Y., B. Asselain, V.J. Bardou, A. Fourquet, T. Hoang, A. Rochefediere and A.D. Tsodikov, 1993. A Simple Stochastic Model of Tumor Recurrence and its Applications to Data on Pre-Menopausal Breast Cancer. In: Biometrics and Analysis Dormees Spatio-Temporal, Asselain, B., M. Boniface, C. Duby, C. Lopez, J.P. Masson and J. Tranchefort (Eds.). French Society of Biometrics, France, pp: 66-82.

Yu, B. and Y. Peng, 2008. Mixture cure models for multivariate survival data. Comput. Stat. Data Anal., 52: 1524-1532.