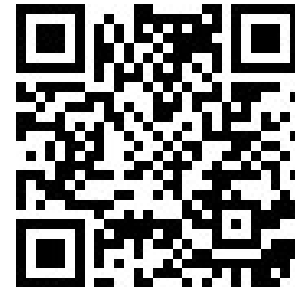


Nadarajah-Haghighi Model for Survival Data With Long Term Survivors in the Presence of Right Censored Data

Umar Usman¹, Suleiman Shamsuddeen²,
Bello Magaji Arkilla³, Yakubu Aliyu^{4*}



*Corresponding author

1. Department of Mathematics, Usmanu Dan Fodiyo University Sokoto-Nigeria, uusman07@gmail.com
2. Department of Mathematics, Usmanu Dan Fodiyo University Sokoto-Nigeria, macostamkd@gmail.com
3. Department of Public Medicine, Usmanu Dan Fodiyo University Sokoto-Nigeria, mbarkilla@gmail.com
4. Department of Mathematics, Usmanu Dan Fodiyo University Sokoto-Nigeria
& Department of Statistics, Ahmadu Bello University Zaria-Nigeria, aliyuyakubu40@yahoo.com

Abstract

In this paper, a new long term survival model called Nadarajah-Haghighi model for survival data with long term survivors was proposed. The model is used in fitting data where the population of interest is a mixture of individuals that are susceptible to the event of interest and individuals that are not susceptible to the event of interest. The statistical properties of the proposed model including quantile function, moments, mean and variance were provided. Maximum likelihood estimation procedure was used to estimate the parameters of the model assuming right censoring. Furthermore, Bayesian method of estimation was also employed in estimating the parameters of the model assuming right censoring. Simulations study was performed in order to ascertain the performances of the *MLE* estimators. Random samples of different sample sizes were generated from the model with some arbitrary values for the parameters for 5%, 1.3% and 1.5% cure fraction values. Bias, standard error and mean square error were used as discrimination criterions. Additionally, we compared the performance of the proposed model with some competing models. The results of the applications indicates that the proposed model is more efficient than the models compared with. Finally, we fitted some models considering type of treatment as a covariate. It was observed that the covariate have effect on the shape parameter of the proposed model.

Key Words: Long term cure rate model, Nadarajah-Haghighi cure rate model, Mixture cure rate mode, Right censoring, Bayesian Estimation.

Mathematical Subject Classification: 60E05, 62E15.

1. Introduction

The mixture cure rate model also known as the standard cure rate model assumes that the population of study is a mixture of two individuals: susceptible individuals who experience the event of interest and non-susceptible (also known as cured or immune) that will never experience the event of interest. Hence, the population of studies consists of individuals that are at risk and individuals that are not at risk (cured group) with respect to the event of interest. To model the proportion of non-susceptible, different approaches; parametric, semi-parametric and non-parametric have been used by different researchers. Interested readers can refer to the work of Boag (1949), Berkson and Gage (1952), Farewell (1986), Meeker (1987), Gamel et al. (1990), Cantor and Shuster (1992), Ng and McLachlan (1998), Peng

et al. (1998), Gieser et al. (1998), Sy and Taylor (2000), Shao and Zhou (2004), Kannan et al. (2010), Cancho et al. (2011), Kutal and Qian (2018) to mention but a few.

According to Kannan et al. (2010), cure rate models have applications in many areas such as health, reliability, criminology and many more areas. For instance, the model have been applied in reliability by Nelsen (2007) while in criminology, a study on recidivism times of prisoners release from prison in western Australia have been reported by Maller and Zhou (1996).

In the present article, we introduced a cure fraction model for survival data based on the Nadarajah-Haghighi distribution in the presence of cure fraction, censoring and covariate. Properties of the model were studied and an application of the model with leukaemia data of Kersey et al. (1987) was provided. The rest of the paper is organized as follows: In section 2, we describe the Nadarajah-Haghighi distribution and provide the survival function, probability density function and cumulative distribution function of the model. Section 3 introduces the Maximum Likelihood Estimation and Bayesian methods of estimation in estimating the parameters of the model assuming right censoring. In section 4, statistical properties of the model such as quantile function and moments of the model were provided. Simulation studies and applications of the model were provided in sections 5 and 6 respectively and finally, we conclude in section 7.

2. Model

2.1. The Nadarajah-Haghighi Distribution

The Nadarajah-Haghighi distribution was introduced by Nadarajah and Haghighi (2011), as an extension of the exponential distribution. The distribution serves as an alternative to the Weibull, generalized exponential and gamma distributions in modeling real life scenarios. Nadarajah-Haghighi exponential distribution reduces to the exponential distribution when the shape parameter takes the value one. The various shapes of the *pdf* and hazard rate function of the distribution for some selected values can be depicted in figures 1 and 2 respectively. The density of this distribution can have decreasing and unimodal shapes while the hazard rate has a decreasing shape, increasing shape and constant. Hence, the shape of the hazard rate function of the *NH* distribution is similar to that of the Generalized exponential distribution. A random variable t with parameters β and λ is said to follow the Nadarajah-Haghighi exponential distribution if its *pdf*, *cdf*, survival function and hazard rate function are respectively given by:

$$f(t/\beta, \lambda) = \beta\lambda(1 + \beta t)^{\lambda-1} \exp\left(1 - (1 + \beta t)^\lambda\right) \quad (1)$$

$$F(t/\beta, \lambda) = 1 - \exp\left(1 - (1 + \beta t)^\lambda\right) \quad (2)$$

$$S(t/\beta, \lambda) = \exp\left(1 - (1 + \beta t)^\lambda\right) \quad (3)$$

and

$$h(t/\beta, \lambda) = \beta\lambda(1 + \beta t)^{\lambda-1} \quad (4)$$

where $\beta > 0$ is the scale parameter and $\lambda > 0$ is the shape parameter.

2.2. The Nagarajah-Haghighi Mixture Cure Rate Model

The mixture cure rate model also referred to as the standard cure rate model was first proposed by Boag (1949). The model assumes that a certain portion p in the population is cured and the remaining portion $1 - p$ is uncured. Hence, the model consists of two components: a component representing the proportion of non-susceptible in the population and a distribution representing the survival experience of the susceptible group. The survival function for the entire population of study is expressed as:

$$S(t) = p + (1 - p) S_u(t) \quad (5)$$

where $0 < p < 1$ is the mixing parameter and $S_u(t)$ is the survival function for the uncured group.

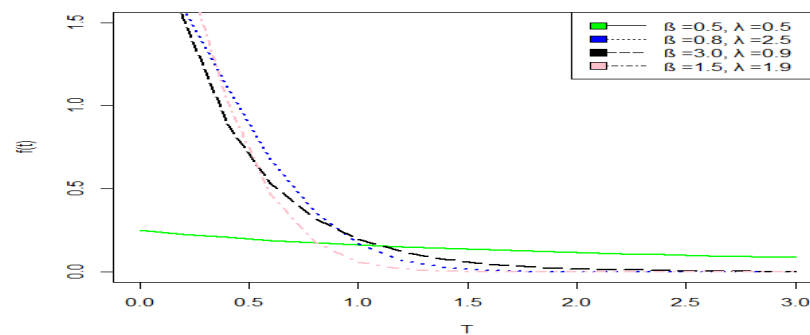


Figure 1: plot of the *pdf* of Nadarajah-Haghighi distribution for some selected values

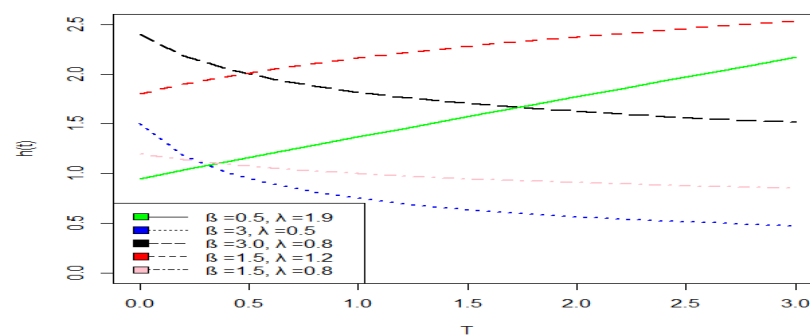


Figure 2: plot of the Hazard function of Nadarajah-Haghighi distribution for some selected values

From the survival function in equation (5), the corresponding probability density function (*pdf*) is given by:

$$f(t) = (1 - p) f_u(t) \quad (6)$$

where $f_u(t)$ is the *pdf* for the uncured individuals.

From equations (2) and (3), the survival function, *pdf* and distribution function of the Nadarajah-Haghighi mixture cure rate model are respectively given by:

$$S(t/\beta, \lambda, p) = p + (1 - p) \exp\left(1 - (1 + \beta t)^\lambda\right) \quad (7)$$

$$f(t/\beta, \lambda, p) = \beta \lambda (1 - p) (1 + \beta t)^{\lambda-1} \exp\left(1 - (1 + \beta t)^\lambda\right) \quad (8)$$

and

$$F(t/\beta, \lambda, p) = (1 - p) \left[1 - \exp\left(1 - (1 + \beta t)^\lambda\right)\right] \quad (9)$$

where $\beta > 0$ is the scale parameter, $\lambda > 0$ is the shape parameter and p is the proportion of uncured individuals.

3. Estimation Procedure

In this section, classical and non-classical methods of estimation were used in estimating the parameters of the Nadarajah-Haghighi mixture cure rate (*NHMC*R) model.

3.1. Maximum Likelihood Estimation Method

Consider a random sample of lifetimes $(t_i, \delta_i, i = 1, 2, \dots, n)$ under the assumption of right censoring, the likelihood function of $(t_i, \delta_i, i = 1, 2, \dots, n)$ is given by:

$$L(\Theta/t, \delta) = \prod_{i=1}^n f(t_i)^{\delta_i} [S(t_i)]^{1-\delta_i} \quad (10)$$

where $\Theta = (\beta, \lambda p)$, δ_i is a censoring indicator, $\delta_i = 1$ if the observed lifetime t_i is uncensored and $\delta_i = 0$ if the observed lifetime is censored. Substituting equations (7) and (8) in equation (10) and taking natural logarithm gives the observed full log-likelihood function as:

$$\ell(\Theta/t, \delta) = z \log(\beta) + z \log(\lambda) + z \log(1-p) + \beta(\lambda-1) \sum_{i=1}^n A_i + \lambda z - \sum_{i=1}^n \delta_i B_i + \sum_{i=1}^n (1-\delta_i) \log[p + (1-p)C_i] \quad (11)$$

where $z = \sum_{i=1}^n \delta_i$, $A_i = \delta_i t_i$, $B_i = (1 + \beta t_i)^\lambda$ and $C_i = \exp(1 - B_i)$. To obtain the score function, we differentiate the log-likelihood function partially with respect to β , λ and p . This gives:

$$\frac{\partial \ell(\Theta/t, \delta)}{\partial \beta} = \frac{z}{\beta} + (\lambda-1) \sum_{i=1}^n A_i - \lambda \sum_{i=1}^n \delta_i D_i t_i - \lambda(1-p) \sum_{i=1}^n \frac{(1-\delta_i) C_i D_i t_i}{p + (1-p) C_i} \quad (12)$$

$$\frac{\partial \ell(\Theta/t, \delta)}{\partial \lambda} = \frac{z}{\lambda} + \beta \sum_{i=1}^n A_i + z - \frac{1}{\lambda} \sum_{i=1}^n \delta_i B_i \log(B_i) - \frac{1-p}{\lambda} \sum_{i=1}^n \frac{(1-\delta_i) B_i \log(B_i) C_i}{p + (1-p) C_i} \quad (13)$$

and

$$\frac{\partial \ell(\Theta/t, \delta)}{\partial p} = -\frac{z}{1-p} + \sum_{i=1}^n \frac{(1-\delta_i)(1-C_i)}{p + (1-p) C_i} \quad (14)$$

where $D_i = (1 + \beta t_i)^{\lambda-1}$. If the log-likelihood function has a global maximizer, then the *MLE* is the solution of the equations in (12), (13) and (14). However, the system of equations in (12) to (14) are non-linear, hence, numerical methods such as Newton-Raphson method are used in obtaining the *MLE*. To obtain interval estimates and test for hypothesis, the observed information matrix $J(\Phi)$ is obtained. The elements of the observed information matrix $J(\Phi)$ are given by:

$$J(\Phi) = - \begin{pmatrix} V_{\lambda\lambda} & V_{\lambda\beta} & V_{\lambda p} \\ & V_{\beta\beta} & V_{\beta p} \\ & & V_{pp} \end{pmatrix}$$

where the diagonal elements of $J(\Phi)^{-1}$ are the variances of the parameters λ , β and p respectively, while the off diagonal elements are the covariances. The elements of $J(\Phi)$ are:

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \lambda^2} &= -\frac{z}{\lambda^2} - \frac{1}{\lambda^2} \sum \delta_i B_i \log^2(B_i) - \frac{1-p}{\lambda^2} \sum \frac{(1-\delta_i) B_i C_i \log^2(B_i)}{p + (1-p) C_i} \left[1 - \frac{B_i p}{p + (1-p) C_i} \right] \\ \frac{\partial^2 \ell}{\partial \beta^2} &= -\frac{z}{\beta^2} - \lambda(\lambda-1) \sum \delta_i t_i^2 B_i^{\frac{\lambda-2}{\lambda}} - \lambda(1-p) \sum \frac{(1-\delta_i) t_i^2 C_i}{p + (1-p) C_i} \\ &\quad \left[(\lambda-1) B_i^{\frac{\lambda-2}{\lambda}} - \frac{p D_i^2 \lambda}{p + (1-p) C_i} \right] \\ \frac{\partial^2 \ell}{\partial p^2} &= -\frac{z}{(1-p)^2} - \sum \frac{(1-\delta_i)(1-C_i)^2}{(p + (1-p) C_i)^2} \\ \frac{\partial^2 \ell}{\partial \beta \partial \lambda} &= \sum \delta_i t_i - \sum \delta_i D_i t_i - \sum \delta_i D_i t_i \log(B_i) - (1-p) \sum \frac{(1-\delta_i) C_i D_i t_i}{(p + (1-p) C_i)} \\ &\quad - (1-p) \sum \frac{(1-\delta_i) C_i D_i \log(B_i)}{(p + (1-p) C_i)} \left[1 - \frac{B_i p}{p + (1-p) C_i} \right] \\ \frac{\partial^2 \ell}{\partial p \partial \lambda} &= \frac{1}{\lambda} \sum \frac{(1-\delta_i) B_i C_i \log(B_i)}{(p + (1-p) C_i)^2} \end{aligned}$$

and

$$\frac{\partial^2 \ell}{\partial p \partial \beta} = \lambda \sum \frac{(1-\delta_i) C_i D_i t_i}{(p + (1-p) C_i)^2}$$

It is important to note that, the asymptotic distribution of $\sqrt{n}(\hat{\Phi} - \Phi)$ is multivariate normal $N_3\left(0, J(\hat{\Phi})^{-1}\right)$, where $J(\hat{\Phi})$ is the total observed information matrix evaluated at $\hat{\Phi}$. The asymptotic $100(1 - \gamma)\%$ confidence interval for the parameters β , λ and p are $\hat{\beta} \pm Z_{\frac{\gamma}{2}} \sqrt{\text{var}(\hat{\beta})}$, $\hat{\lambda} \pm Z_{\frac{\gamma}{2}} \sqrt{\text{var}(\hat{\lambda})}$ and $\hat{p} \pm Z_{\frac{\gamma}{2}} \sqrt{\text{var}(\hat{p})}$ respectively, where $Z_{\frac{\gamma}{2}}$ is the $100(1 - \gamma)\%$ quantile of the standard normal distribution.

In the presence of a vector of covariates $\mathbf{x} = (x_1, x_2, \dots, x_m)$ affecting the parameters β , λ and p , we assumed the regression model $\beta_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi})$, $\lambda_i = \exp(\lambda_0 + \lambda_1 x_{1i} + \lambda_2 x_{2i} + \dots + \lambda_m x_{mi})$ and $\log\left(\frac{p_i}{1-p_i}\right) = \eta_0 + \eta_1 x_{1i} + \eta_2 x_{2i} + \dots + \eta_m x_{mi}$. For instance, assuming one covariate x_i for $i = 1, 2, \dots, n$ affecting the parameter β , λ and p , the link function for the parameters becomes $\beta_i = \exp(\beta_0 + \beta_1 x_i)$, $\lambda_i = \exp(\lambda_0 + \lambda_1 x_i)$ and $\log\left(\frac{p_i}{1-p_i}\right) = \eta_0 + \eta_1 x_i$, where x_i takes the value 0 when individual i is in the treatment group 1 and the value 1 when individual i is in the treatment group 2.

3.2. Bayesian Method of Estimation

Consider the *NHMCR* model proposed in sections 2.2, let Θ be the vector of unknown parameters. The joint distribution of the model parameters is obtain by combining the joint prior distribution of the parameters and the likelihood function. A gamma prior for the shape and scale parameters of the proposed models are assumed, where *Gamma* (r, q) denotes gamma distribution with mean $\frac{r}{q}$ and variance $\frac{r}{q^2}$, r and q are hyper-parameters. On the other hand, the proportion of cure parameter is assumed to follow the beta prior. That is, $p \sim \text{Beta}(s, v)$, where s and v are hyper-parameters. To be specific, $p \sim \text{Beta}(1, 1)$ since $0 < p < 1$. We further assume independent prior among the parameters of the model.

In the presence of m covariates $\mathbf{x} = (x_1, x_2, \dots, x_m)$ affecting the parameters of the proposed models, a link function for the parameters β , λ and p is assumed. That is,

$$\begin{aligned}\log(\beta) &= \beta_0 + \beta_1 x_i + \dots + \beta_m x_m \\ \log(\lambda) &= \lambda_0 + \lambda_1 x_i + \dots + \lambda_m x_m\end{aligned}$$

and

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_0 + \eta_1 x_i + \dots + \eta_m x_m$$

for β , λ and p respectively. To be specific, assume the *NHMCR* model with scale parameter β , shape parameter λ and a proportion of cure p are affected by the presence of a covariate x_i for $i = 1, 2, \dots, n$, then the link function

$$\begin{aligned}\log(\beta) &= \beta_0 + \beta_1 x_{1i} \\ \log(\lambda) &= \lambda_0 + \lambda_1 x_{1i}\end{aligned}$$

and

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_0 + \eta_1 x_{1i}$$

are assumed for the scale parameter β , shape parameter λ and proportion of cure p respectively. Inferences considering covariate effect are obtain by replacing β , λ and p with these link functions. Furthermore, normal prior distribution will be assumed for the effect of covariates. That is, $N(\mu, \sigma^2)$ will be assumed, where $N(\mu, \sigma^2)$ is normal distribution with parameters μ and σ .

4. Statistical Properties of the Model

In this section, some statistical properties such as quantile function, median and moments of the *NHMCR* model were discussed.

4.1. Quantile Function and Simulations

In simulation studies, quantile function is used in obtaining random realizations from a given model. The quantile function of the *NHMCR* model is:

$$Q(u) = \frac{1}{\beta} \left[\left[1 - \log \left(\frac{1-p-u}{1-p} \right) \right]^{\frac{1}{\lambda}} - 1 \right] \quad (15)$$

where u is a random number generated from uniform distribution with parameters zero and one. That is $u \sim U(0, 1)$. The quantile function can be use in obtaining the first, second and third quantiles of the *NHMCR* model. This is done by letting $u = \frac{1}{4}, \frac{1}{2}$ and $\frac{3}{4}$ respectively for the first, second and third quantiles. For instance, to obtain the median, let $u = \frac{1}{2}$, we obtain the median of the *NHMCR* model as:

$$Median = \frac{1}{\beta} \left[\left[1 - \log \left(\frac{0.5-p}{1-p} \right) \right]^{\frac{1}{\lambda}} - 1 \right] \quad (16)$$

To simulate a random sample of size n from the *NHMCR* model with right censored data, the following algorithm is followed:

1. Generate a random sample $u_i, i = 1, 2, \dots, n$ from $U(0, 1)$.
2. For a cure fraction p , return $t_i = \frac{1}{\beta} \left[\left[1 - \log \left(\frac{1-p-u}{1-p} \right) \right]^{\frac{1}{\lambda}} - 1 \right]$ for $u_i < 1-p$ otherwise t_i is infinity.
3. Generate a sample of the censoring times $c_i, i = 1, 2, \dots, n$ from the *NH* distribution.
4. Calculate $z_i = \min(t_i, c_i), \delta_i = I(t_i \leq c_i), i = 1, 2, \dots, n$
5. The observed data set $D = \{(z_i, \delta_i), i = 1, 2, \dots, n\}$ are realizations from the *NHMCR* model with right censoring.

4.2. Moments of the Model

Let T be a random variable that follows the *NHMCR* with *pdf* given by (8). The r th moment about the origin of the *NHMCR* model is obtain as follows:

$$E(T^r) = \int_0^\infty t^r \beta \lambda (1-p) (1+\beta t)^{\lambda-1} \exp \left(1 - (1+\beta t)^\lambda \right) dt$$

substituting $m = (1+\beta t)^\lambda$ the integral can be written as:

$$\begin{aligned} E(T^r) &= \frac{(1-p)e}{\beta^r} \int_1^\infty \left(m^{\frac{1}{\lambda}} - 1 \right)^r e^{-m} dm \\ &= \frac{(1-p)e}{\beta^r} \sum_{j=0}^r (-1)^{r+j} {}^r C_j \Gamma \left(\frac{j}{\lambda} + 1, 1 \right) \end{aligned}$$

where $\Gamma(p, b) = \int_b^\infty t^{p-1} e^{-t} dt$ is the complementary incomplete gamma function. Hence, the first four moments of the model are:

$$E(T) = \frac{(1-p)}{\beta} \left[-1 + e \Gamma \left(\frac{1}{\lambda} + 1, 1 \right) \right] \quad (17)$$

$$E(T^2) = \frac{(1-p)}{\beta^2} \left[1 - 2e \Gamma \left(\frac{1}{\lambda} + 1, 1 \right) + e \Gamma \left(\frac{2}{\lambda} + 1, 1 \right) \right] \quad (18)$$

$$E(T^3) = \frac{(1-p)}{\beta^3} \left[-1 + 3e \Gamma \left(\frac{1}{\lambda} + 1, 1 \right) - 3e \Gamma \left(\frac{2}{\lambda} + 1, 1 \right) + e \Gamma \left(\frac{3}{\lambda} + 1, 1 \right) \right] \quad (19)$$

and

$$E(T^4) = \frac{(1-p)}{\beta^4} \left[1 - 4e\Gamma\left(\frac{1}{\lambda} + 1, 1\right) + 6e\Gamma\left(\frac{2}{\lambda} + 1, 1\right) - 4e\Gamma\left(\frac{3}{\lambda} + 1, 1\right) + e\Gamma\left(\frac{4}{\lambda} + 1, 1\right) \right] \quad (20)$$

Hence, the variance of the *NHMC*R model is obtain by $\sigma^2 = E(T^2) - (E(T))^2$ which is given by:

$$\sigma^2 = \frac{(1-p)}{\beta^2} \left[e\Gamma\left(\frac{2}{\lambda} + 1, 1\right) - (1-p)e^2\Gamma\left(\frac{1}{\lambda} + 1, 1\right) - 2e\Gamma\left(\frac{1}{\lambda} + 1, 1\right) + p \right]$$

Table 1: Summary statistics with 5% cure fraction

n	parameters	estimates	bias	Standard Error	MSE
$\beta = 1.5$ and $\lambda=2.0$					
100	β	1.5328	0.0328	1.0430	1.8950
	λ	2.7259	0.7259	1.4826	4.3504
	p	0.0465	-0.0035	0.02243	0.0054
200	β	1.4986	-0.0141	0.7671	0.9171
	λ	2.4076	0.4076	0.9906	1.6069
	p	0.0497	-0.0003	0.0155	0.0005
300	β	1.4870	-0.0130	0.6130	0.1134
	λ	2.2800	0.2800	0.7706	1.1499
	p	0.0500	-0.0002	0.0127	0.0003
400	β	1.4792	-0.0121	0.5444	0.0249
	λ	2.2230	0.2230	0.6649	1.0416
	p	0.0500	0.0000	0.0110	0.0003
500	β	1.4877	-0.0120	0.4709	0.3299
	λ	2.1613	0.1613	0.5627	0.5713
	p	0.0500	0.0000	0.0099	0.0002
$\beta = 2.5$ and $\lambda = 2.0$					
100	β	2.4433	-0.0568	1.7541	6.6157
	λ	2.7825	0.7825	1.5725	5.1135
	p	0.0495	-0.0005	0.0219	0.0011
200	β	2.4724	-0.0476	1.2108	2.4757
	λ	2.4003	0.4003	0.9624	2.2915
	p	0.0498	-0.0002	0.0153	0.0005
300	β	2.4551	-0.0449	0.9881	1.7599
	λ	2.2790	0.2790	0.7803	1.6454
	p	0.0499	-0.0001	0.0125	0.0003
400	β	2.4522	-0.0378	0.8342	0.9515
	λ	2.2197	0.21971	0.6082	0.5942
	p	0.0500	-0.0000	0.0108	0.0002
500	β	2.4641	-0.0359	0.7539	0.1950
	λ	2.1624	0.1624	0.5736	0.1581
	p	0.0500	0.0000	0.0097	0.0002

5. Simulation Study

In this section, simulation studies was conducted in order to ascertain the performance of the maximum likelihood estimator of Θ (where $\Theta = (\beta, \lambda, p)'$) discussed in section 3. The algorithm discussed in section 4.1 was used in generating right censored survival times.

We generated random samples of size $n = 100, 200, 300, 400$ and 500 from the *NHMC*R model with some arbitrary parameters $\beta = 1.5, \lambda = 2.0$ and $\beta = 2.5, \lambda = 2.0$ with different values for cure fraction. On the other hand, the censoring variable follows the *NH* distribution. Inference results from the simulation study were based on 1000 replications. These results were obtained under the *MLE* estimators. Performance measures such as bias, standard error and mean square error were employed so as to ascertain the performances of the estimates.

Table 2: Summary statistics with 10% cure fraction

n	parameters	estimates	bias	SE	MSE
$\beta=1.5, \lambda=2.0$					
100	β	1.4743	-0.0357	1.0196	2.1285
	λ	2.7259	1.5986	2.4423	6.6399
	p	0.0969	-0.0014	0.0299	0.0018
200	β	1.489	-0.031	0.6904	0.9865
	λ	2.4194	0.7194	1.1944	3.902
	p	0.0994	-0.0006	0.0212	0.0009
300	β	1.473	-0.027	0.5541	0.6425
	λ	2.4026	0.4026	0.884	0.7631
	p	0.0994	-0.0006	0.0174	0.0006
400	β	1.4776	-0.0224	0.476	0.4583
	λ	2.2356	0.2356	0.6266	0.5307
	p	0.0996	-0.0004	0.0151	0.0005
500	β	1.4796	-0.0204	0.4242	0.369
	λ	2.1742	0.1742	0.5405	0.4537
	p	0.0997	-0.0003	0.0135	0.0004
$\beta=2.5, \lambda=2.0$					
100	β	2.4396	-0.0694	1.5922	5.1052
	λ	2.8135	1.6105	1.0748	6.8271
	p	0.0996	-0.0004	0.0295	0.0018
200	β	2.4522	-0.0678	1.0997	2.4709
	λ	2.6405	0.6405	1.0643	4.2042
	p	0.0996	-0.0004	0.021	0.0009
300	β	2.4387	-0.0613	0.8933	1.5967
	λ	2.3451	0.3451	0.8651	2.1335
	p	0.0995	-0.0005	0.0172	0.0006
400	β	2.4517	-0.0483	0.7659	1.2055
	λ	2.226	0.226	0.6218	1.1435
	p	0.0996	-0.0004	0.0149	0.0004
500	β	2.4565	-0.0435	0.6801	0.954
	λ	2.1713	0.1713	0.5139	0.7139
	p	0.0998	-0.0002	0.0134	0.0004

Tables 1 to 3 gives the simulation results based on 1000 replications for each parameter setting. The Table gives the mean estimate of the parameters together with the bias, standard error and mean square error (*MSE*). It is observed that, the estimates are satisfactory closed to the true parameter values in all the different simulation settings. It is also observed that the estimates of bias, standard error and mean square error (*MSE*) decreases as sample size increases for all the different settings of all examined parameter values. Hence, the results reveal that the proposed model has a good performance.

6. Real data Analysis

In this section, a data set consisting of 90 observations that was provided by Kersey et al. (1987) and was analyzed by Shao and Zhou (2004), Coelho-Barros et al. (2017) and Kutal and Qian (2018) were use to fit the mixture cure rate model studied in previous sections and compared its performance with that of some competing mixture models. The data recorded the times to recurrence of leukaemia for patients after one of allogeneic transplant or autologous transplant. Forty-six (46) patients were treated by allogeneic transplant while the remaining forty-four (44) were treated by autologous transplant. However, there are thirty-three (33) and thirty-five patients treated by allogeneic transplant and autologous transplant respectively that suffered a recurrence of leukaemia at different ranges of time. It is also observed that thirteen(13) and nine(9) patients from these respective groups have no record of recurrence, that is they are censored. Maximum Likelihood method of estimation and Bayesian method of estimation discussed in section 3 were used to analyzed this data.

The data is fitted in three categories: Models without cure fraction assuming right censoring, models assuming cure

fraction under right censoring and models assuming cure fraction and covariates under right censoring. In the first category, Nadarajah-Haghighi(NH), Weibull (W), Generalized Gompertz (GG) and Rayleigh (R) distributions under right censoring were fitted.

Table 3: Summary statistics with 1.3% cure fraction

n	parameters	estimates	bias	SE	MSE
$\beta=1.5, \lambda=2.0$					
100	β	1.4602	-0.098	1.0164	2.0734
	λ	2.9438	1.8431	1.9309	7.6266
	p	0.129	-0.001	0.0335	0.0023
200	β	1.4632	-0.068	0.7006	1.0187
	λ	2.8276	0.8276	1.1365	5.2447
	p	0.1293	-0.0007	0.0238	0.0011
300	β	1.4755	-0.0245	0.5621	0.677
	λ	2.434	0.434	0.8753	4.9105
	p	0.1295	-0.0005	0.0195	0.0008
400	β	1.4782	-0.0218	0.485	0.4794
	λ	2.2247	0.2247	0.6583	1.1045
	p	0.1298	-0.0002	0.0169	0.0006
500	β	1.4823	-0.0177	0.4326	0.3902
	λ	2.1857	0.1857	0.5508	0.8913
	p	0.1299	-0.0001	0.0151	0.0005
$\beta=2.5, \lambda=2.0$					
100	β	2.433	-0.067	1.606	5.1599
	λ	2.7338	1.0938	1.9384	6.1309
	p	0.1294	-0.0008	0.0332	0.0023
200	β	2.4679	-0.0521	1.1284	2.6396
	λ	2.6654	0.6654	1.1176	4.7618
	p	0.1293	-0.0007	0.0236	0.0011
300	β	2.4509	-0.0491	0.9024	1.7106
	λ	2.3762	0.3762	0.7926	3.0792
	p	0.1296	-0.0004	0.0193	0.0007
400	β	2.4511	-0.0489	0.7744	1.2442
	λ	2.2221	0.2221	0.6208	1.0741
	p	0.1299	-0.0001	0.0168	0.0006
500	β	2.4663	-0.0337	0.6954	1.0081
	λ	2.168	0.168	0.5275	0.8044
	p	0.1298	-0.0002	0.015	0.0004

Then, the *NHMCR* model was fitted and compared with the fits of Weibull (*WMCR*), Generalized Gompertz (*GGMCR*) and Rayleigh (*RMCR*) mixture cure rate models. Likelihood ratio test was then performed to ascertain the importance of the addition of the cure fraction parameter.

Table 4: Maximum Likelihood Estimates (MLE) and Standard Error for models without cure fraction

Model	parameters	estimate	SE	95% CI	AIC	BIC	CAIC
Nadarajah-Haghighi	β	0.2997	0.0437	(0.2141, 0.3853)	160.6106	160.5191	156.7485
	λ	6.8500	2.3940	(2.1578, 4.9919)			
Weibull	β	0.7957	0.0981	(0.6034, 0.9880)	178.8961	178.8046	175.034
	λ	0.6441	0.0626	(0.4519, 0.8364)			
Generalized Gompertz	β	0.6581	0.0431	(0.1133, 0.2075)	669.2243	669.087	663.5034
	λ	0.1604	0.0240	(0.5737, 0.7425)			
	θ	1.1484	0.0330	(1.0838, 1.2130)			
Rayleigh	β	2.3330	0.2809	(1.7825, 2.8835)	412.4524	412.4066	410.4978

In the last part of the analysis, the *NHMCR* model was fitted in the presence of a covariate on the parameters of the

model.

6.1. Maximum Likelihood Result

In this section, the results of the fitted models under *MLE* procedure is discussed. Different performance measures such as Akaike information criteria (*AIC*), Bayesian information criteria (*BIC*) and Consistent Akaike information criteria (*CAIC*) were used in finding the best fitted model. The model with the least statistic values of these criteria is considered to be the best model.

Tables 4 and 5 gives the *MLE* results of the models without cure fraction and the models with cure fraction respectively. These tables also contain the *AIC*, *BIC* and *CAIC* values of the fitted models. These results reveal that, all the estimates of the models with cure fraction fits the data more efficient than the models without cure fraction.

Table 5: Maximum Likelihood Estimates (MLE) and Standard Error for models with cure fraction

Model	parameters	estimate	SE	95% CI	AIC	BIC	CAIC
NHMCR	β	2.5125	0.9977	(0.5571, 4.4679)	146.4447	146.3074	140.7237
	λ	0.8454	0.2001	(0.4533, 2.9046)			
	p	0.2268	0.0452	(0.1382, 0.3153)			
WMCR	β	1.7214	0.2324	(1.2660, 2.1768)	171.6146	171.4773	165.8937
	λ	1.1169	0.0968	(0.9272, 1.3066)			
	p	0.3756	0.0452	(0.2870, 0.4641)			
GGMCR	β	22.6825	1.0486	(141.8530, 142.0261)	198.6840	198.5010	191.1546
	λ	141.9396	0.0442	(20.6273, 24.7376)			
	θ	19.8721	0.7428	(18.4163, 21.3279)			
	p	0.2296	0.0442	(0.1430, 0.3161)			
RMCR	β	0.5032	0.0304	(0.4436, 0.5628)	205.5954	205.5039	201.7333
	p	0.2332	0.0446	(0.1458, 0.3206)			

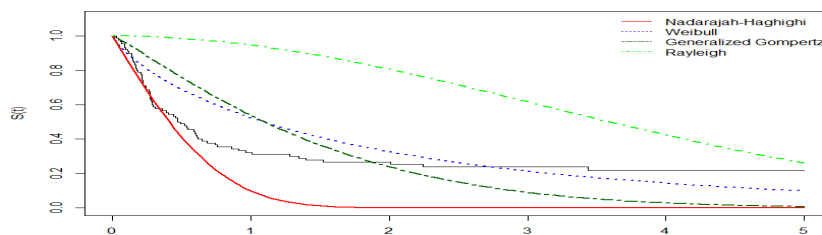


Figure 3: plot of Kaplan-Meier together with NH, GG, W and R models from *MLE* results

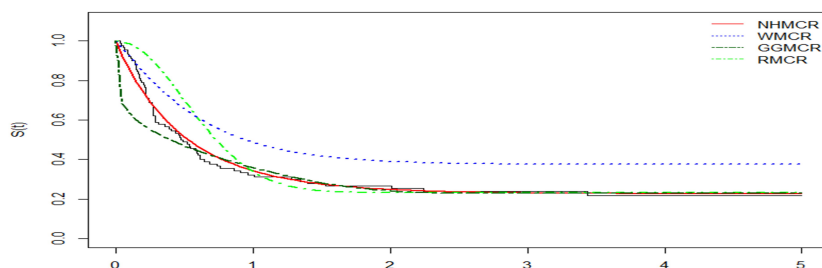


Figure 4: plot of Kaplan-Meier with NHMCR, GGMCR, WMCR and RMCR models from *MLE* results

Hence, the mixture cure rate models are better than the models without cure fraction parameter. This can clearly be seen in figures 3 and 4. Likelihood ratio test was conducted to ascertain the significance of the addition of the cure

fraction parameter in the NH , W , GG and R distributions. The test was conducted at 5% level of significance, we formulate the hypothesis $H_0 : p = 0$ against $H_1 : p \neq 0$. The LRT statistic and p -value for each test is given in table 6, from this table, we observed that the cure fraction parameter is significant in modeling this data.

Table 6: Likelihood Ratio Test

Model compared	τ	p value
NHMCR vs NHD	16.165900	0.000058
WMCR vs WD	9.281480	0.002315
GGMCR vs GGD	472.540300	0.000000
RMCR vs RD	208.857000	0.000000

On the other hand, comparing between the fitted mixture cure rate models ($NHMCR$, $WMCR$, $GGMCR$ and $RMCR$), we conclude that the proposed $NHMCR$ model is more efficient than $WMCR$, $GGMCR$ and $RMCR$ models since it has the least values for AIC , BIC and $CAIC$.

6.2. Bayesian Result

In this section, Bayesian method of estimation was used in fitting the model proposed in the previous section and compared its performance with that of the aforementioned models in section 6.1.

Table 7: Posterior summaries of models not including cure fraction

Model	parameter	posterior mean	SD	95% credible interval	DIC	EBIC
Nadarajah	β	2.4120	0.7693	(0.0932, 4.2180)	206.3	208.6
Haghighi	λ	0.5858	0.0932	(0.4361, 0.8004)		
Weibull	β	0.9439	0.1052	(0.7489, 1.1610)	210.9	212.8
	λ	0.8210	0.0653	(0.6979, 0.9535)		
Generalized	β	0.0629	0.0551	(0.0019, 0.2022)	215.0	218.8
Gompertz	λ	0.6073	0.1170	(0.3774, 0.8389)		
	θ	0.7239	0.1036	(0.5344, 0.9410)		
Rayleigh	β	0.2821	0.0297	(0.2266, 0.3426)	391.8	392.8

In analyzing the data, the prior distribution for the shape and scale parameters were assumed to be Gamma(1, 1) while the prior distribution for the cure fraction parameter was assumed to be Beta(1, 1). Additionally, in the presence of covariates, a normal prior was assumed ($N(0, 100)$) for the covariate effects.

Table 8: Posterior summaries assuming the mixture models

Model	parameter	posterior mean	SD	95% credible interval	DIC	EBIC
Nadarajah-	β	2.2110	0.7851	(1.0250, 4.0550)	145.9	149.2
Haghighi	λ	0.9542	0.2194	(0.6136, 1.4580)		
	p	0.2309	0.0454	(0.1485, 0.3249)		
Generalized	β	0.0748	0.0678	(0.0022, 0.2528)	146.3	151.0
Gompertz	λ	2.0790	0.3629	(1.4000, 2.8130)		
	θ	1.2190	0.2142	(0.8412, 1.6870)		
	p	0.2360	0.0441	(0.1538, 0.3271)		
Weibull	β	1.9110	0.2476	(1.4490, 2.4270)	146.6	149.5
	λ	1.0460	0.0949	(0.8649, 1.2350)		
	p	0.2339	0.0448	(0.1516, 0.3265)		
Rayleigh	β	1.7070	0.2056	(1.6990, 2.1320)	223.9	225.8
	p	0.2390	0.0441	(0.2373, 0.3305)		

Inference results of interest were obtained from simulated samples for the joint posterior distribution using Markov Chain Monte Carlo simulations (MCMC) technique. In all the fitted models, we generated 1010000 samples for each parameter of interest. To minimize the effect of initial values, the first 10000 samples were discarded as burn-in-samples. Furthermore, the posterior summaries of interest were based on 20000 samples taken from the 1000000 samples taken at every 50th sample so as to have approximately uncorrelated values.

To compare between the fitted models under the Bayesian frame work, deviance information criteria (*DIC*) and expected Bayesian information criteria (*EBIC*) were used.

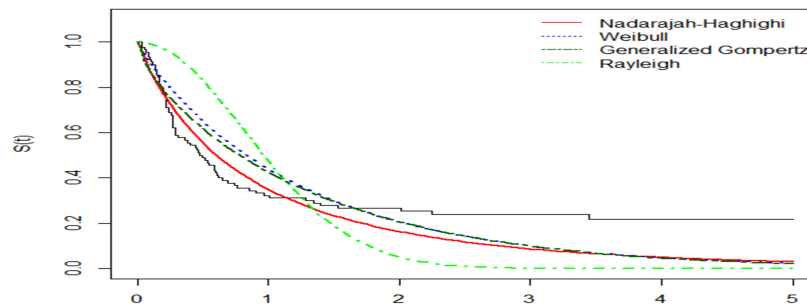


Figure 5: plot of Kaplan-Meier with NH, GG, W and R models from Bayesian results

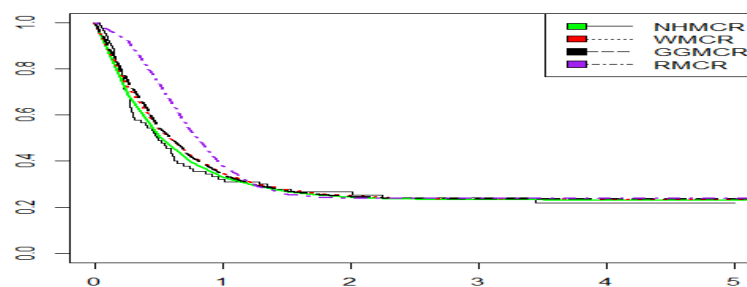


Figure 6: plot of Kaplan-Meier with NHMCR, GGMCR, WMCR and RMCRC models from Bayesian results

Table 7 gives the fits of *NH*, *GG*, *W* and *R* distributions without cure fraction. The results gives the posterior mean, standard deviation (SD), 95% credible interval and the statistics: *DIC* and *EBIC*. Figure 5 gives the Kaplan-Meier curve together with that of *NH*, *W*, *GG* and *R* distributions. Similar to the fits in figure 3, these fits does not properly mimic that of the empirical survival function. Additionally, the presence of plateau close to 0.3 suggest cure fraction models as an alternatives to these fitted models. Table 8 gives the posterior summaries of the *NHMCR*, *GGMCR*, *WMCR* and *RMCR* models. It can easily be seen that the cure rate models fits the data better than the models without cure fraction. It is also observed that our proposed *NHMCR* model is a very strong competitor since it has the least values of *DIC* and *EBIC*. This can also be observed from the Kaplan-Meier fits together with that of the fitted mixture models as shown in figure 6.

6.3. Covariate Models

In this section, we fit the *NHMCR* model considering the type of treatment as a covariate. For the regression model given in section 3, six different models were considered: model I - regression model for β , model II - regression model for λ , model III - regression model for p , model IV - regression model for β and λ , model V - regression model for λ and p and model VI - regression model for β , λ and p .

The results of the regression models considering the *NHMCR* model under the *MLE* are given in table 9 while the estimates under the Bayesian technique are given in table 11. The information criteria for these fitted models are given in table 10. Careful examination of the information criteria values shows that, the model considering covariate on the shape parameter is more efficient since it has the least estimates of *AIC*, *BIC*, *DIC* and *EBIC* values.

Table 9: Maximum Likelihood Estimates (MLE) and Standard Error for regression models

Models	Parameters	estimate	SE	95% CI
Model I	β_0	0.1743	0.4914	(-0.7889, 1.1376)
	β_1	0.7083	0.2474	(0.2234, 1.1932)
	λ	1.1126	0.3296	(0.4666, 1.7586)
	p	0.2284	0.0454	(0.1394, 0.3174)
Model II	β	1.9033	0.8957	(0.1478, 3.6588)
	λ_0	-0.1831	0.3067	(-0.7842, 0.4180)
	λ_1	0.4466	0.1669	(0.1194, 0.7737)
	p	0.2296	0.0464	(0.1386, 0.3205)
Model III	β	2.3431	1.0947	(0.1976, 4.4886)
	λ	0.8918	0.2478	(0.4062, 1.3774)
	η_0	-0.9400	0.3321	(-1.5908, -0.2892)
	η_1	-0.6102	0.5234	(-1.6361, 0.4157)
Model IV	β_0	1.0334	0.7331	(-0.4035, 2.4703)
	β_1	-0.5978	0.9204	(-2.4017, 1.2061)
	λ_0	-0.4718	0.4771	(-1.4070, 0.4633)
	λ_1	0.8813	0.6073	(-0.3090, 2.0715)
	p	0.2276	0.0511	(0.1275, 0.3277)
Model V	β	2.1835	0.8156	(0.5850, 3.7821)
	λ_0	-0.2828	0.257	(-0.7864, 0.2209)
	λ_1	0.4327	0.1758	(0.0881, 0.7773)
	η_0	-1.0088	0.3572	(-1.7088, -0.3087)
	η_1	-0.4719	0.5273	(-1.5053, 0.5615)
Model VI	β_0	1.2340	0.7779	(-0.2906, 2.7586)
	β_1	-1.0183	0.9791	(-2.9372, 0.9006)
	λ_0	-0.4622	0.4357	(-1.3162, 0.3918)
	λ_1	1.0072	0.6016	(-0.1719, 2.1864)
	η_0	-1.1675	0.3845	(-1.9212, -0.4138)
	η_1	-0.3040	0.5461	(-1.3743, 0.7662)

Table 10: Information criterion

models	AIC	BIC	CAIC	DIC	EBIC
I	142.1614	141.9784	134.6320	145.9000	151.5170
II	140.6567	140.4737	133.1273	141.9000	147.4170
III	147.2024	147.0194	139.6730	161.2000	185.5674
IV	142.3348	142.1061	133.0491	142.4000	150.3000
V	141.8601	141.6313	132.5743	157.3000	165.4712
VI	144.3350	144.0605	133.3471	157.8000	168.2255

7. Conclusion

The presence of cure fraction and covariates usually occur in lifetime data analysis especially medical applications. To analyze this type of data, the cure fraction model is used. This paper proposed a mixture cure fraction model with *NH* susceptible distribution for right censored data. Results of simulation study shows that, the proposed model has a good performance. Moreover, a real data was used in comparing the performance of the proposed model with *WMCR*, *GGMCR* and *RMCR* models. It was found that the proposed model fits the data more efficient than the aforementioned models. Finally, we analyzed the data assuming type of treatment as a covariate. Six different models were fitted and it was found that model *II* was better fitted to the data under the *MLE* and Bayesian method of estimations.

Table 11: Posterior summaries assuming the mixture model and including covariate

model	parameter	posterior mean	SD	95% credible interval
Model I	β_0	0.0413	0.0980	(-0.1493, 0.2340)
	β_1	0.1257	0.0913	(-0.0534, 0.3052)
	λ	1.3860	0.1390	(1.1230, 1.6700)
	p	0.2350	0.0445	(0.1529, 0.3273)
Model II	β	1.8350	0.3537	(1.2420, 2.6240)
	λ_0	-0.0142	0.0912	(-0.1890, 0.1668)
	λ_1	0.1265	0.0866	(-0.0446, 0.2944)
	p	0.2341	0.0445	(0.1532, 0.3277)
Model III	β	2.1240	0.7326	(0.9999, 3.8440)
	λ	0.9917	0.2145	(0.6639, 1.5020)
	η_0	-0.1890	0.0906	(-0.3642, -0.01324)
	η_1	-0.1084	0.0958	(-0.2960, 0.0812)
Model IV	β_0	0.1184	0.0887	(-0.0587, 0.2919)
	β_1	0.1175	0.0930	(-0.0646, 0.3000)
	λ_0	0.1314	0.0770	(-0.0244, 0.2790)
	λ_1	0.1572	0.0878	(-0.0167, 0.3296)
	p	0.2295	0.0451	(0.1467, 0.3234)
Model V	β	1.8520	0.3514	(1.2650, 2.6270)
	λ_0	-0.0080	0.0903	(-0.1839, 0.1702)
	λ_1	0.1261	0.0856	(-0.0442, 0.2931)
	η_0	-0.1882	0.0908	(-0.3653, -0.0089)
	η_1	-0.1087	0.0953	(-0.2947, 0.0801)
Model VI	β_0	0.1230	0.0883	(-0.0498, 0.2961)
	β_1	0.1151	0.0927	(-0.0649, 0.2971)
	λ_0	0.1458	0.0748	(-0.0036, 0.2896)
	λ_1	0.1551	0.0866	(-0.0137, 0.3251)
	η_0	-0.1888	0.0911	(-0.3698, -0.0097)
	η_1	-0.1089	0.0963	(-0.2968, 0.0804)

References

1. Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515.
2. Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53.
3. Cancho, V. G., Rodrigues, J., and de Castro, M. (2011). A flexible model for survival data with a cure rate: a bayesian approach. *Journal of Applied Statistics*, 38(1):57–70.
4. Cantor, A. B. and Shuster, J. J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine*, 11(7):931–937.
5. Coelho-Barros, E. A., Achcar, J. A., and Mazucheli, J. (2017). Cure rate models considering the burr xii distribution in presence of covariate. *Journal of Statistical Theory and Applications*, 16(2):150–164.
6. Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262.
7. 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(8):999–1006.
8. Gieser, P. W., Chang, M. N., Rao, P., Shuster, J. J., and Pullen, J. (1998). Modelling cure rates using the gompertz model with covariate information. *Statistics in medicine*, 17(8):831–839.
9. Kannan, N., Kundu, D., Nair, P., and Tripathi, R. C. (2010). The generalized exponential cure rate model with covariates. *Journal of Applied Statistics*, 37(10):1625–1636.
10. Kersey, J. H., Weisdorf, D., Nesbit, M. E., LeBien, T. W., Woods, W. G., McGlave, P. B., Kim, T., Vallera, D. A., Goldman, A. I., Bostrom, B., et al. (1987). Comparison of autologous and allogeneic bone marrow transplantation

- for treatment of high-risk refractory acute lymphoblastic leukemia. *New England Journal of Medicine*, 317(8):461–467.
11. Kutal, D. and Qian, L. (2018). A non-mixture cure model for right-censored data with frechet distribution. *Stats*, 1(1):176–188.
12. Maller, R. A. and Zhou, X. (1996). *Survival analysis with long-term survivors*. Wiley New York.
13. Meeker, W. Q. (1987). Limited failure population life tests: application to integrated circuit reliability. *Technometrics*, 29(1):51–65.
14. Nadarajah, S. and Haghighi, F. (2011). An extension of the exponential distribution. *Statistics*, 45(6):543–558.
15. Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.
16. Ng, S. and McLachlan, G. (1998). On modifications to the long-term survival mixture model in the presence of competing risks. *Journal of Statistical Computation and Simulation*, 61(1-2):77–96.
17. Peng, Y., Dear, K. B., and Denham, J. (1998). A generalized f mixture model for cure rate estimation. *Statistics in medicine*, 17(8):813–830.
18. Shao, Q. and Zhou, X. (2004). A new parametric model for survival data with long-term survivors. *Statistics in medicine*, 23(22):3525–3543.
19. Sy, J. P. and Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236.