# The generalized exponential cure rate model with covariates

Nandini Kannan[1], Debasis Kundu[2], P. Nair[3] and R.C. Tripathi[1]

Abstract In this article, we consider a parametric survival model that is appropriate when the population of interest contains long-term survivors or *immunes*. The model referred to as the cure rate model was introduced by Boag [1] in terms of a mixture model that included a component representing the proportion of immunes and a distribution representing the life times of the *susceptible* population. We propose a cure rate model based on the generalized exponential distribution that incorporates the effects of risk factors or covariates on the probability of an individual being a long-time survivor.

Maximum likelihood estimators of the model parameters are obtained using the the EM algorithm. A graphical method is also provided for assessing the goodness of fit of the model. We present an example to illustrate the fit of this model to data that examines the effects of different risk factors on relapse time for drug addicts.

**Keywords and Phrases:** Cure rate, long-term survivor, generalized exponential distribution, EM algorithm, goodness of fit.

**Corresponding Author:** Debasis Kundu

[1] Department of Management Science and Statistics, The University of Texas at San Antonio, 1604 N.West Loop, San Antonio, Texas, USA.

[2] Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Pin 208016, INDIA.

[3] Department of Statistics, Rice University, Houston, Texas, USA.

# 1 Introduction

In recent years the development of new drugs and treatment regimens has resulted in patients living longer with diseases such as cancer and heart disease. In cohorts of patients with certain types of cancer, it is observed that some patients are cured permanently, i.e. show no recurrence of the disease. The patients who are cured are called immunes or long-term survivors, while the remaining patients who develop a recurrence of the disease are termed susceptibles. The population of interest may thus be regarded as a mixture of these two types of patients. Standard parametric and non-parametric survival models are inappropriate for analyzing such data because they ignore the distinction between the immunes and susceptibles in the population.

Boag [1] first proposed a two-component mixture model for analyzing breast cancer data. The model he proposed is referred to as the "cure rate" model, and is formulated in terms of a mixture model. The model introduces a component representing the proportion of immunes in the population and a distribution representing the survival experience of the susceptibles, called the latency distribution.

There are many applications of this model in areas such as health, criminology, reliability and economics. For example, in a study of leukemia patients, Freireich *et al.* [6] compared the survival experiences of patients given the drug 6-MP compared to a control group. They noted that all patients in the control group experienced symptoms of the disease during the study, while more than half of the patients in the treatment group did not exhibit any signs of remission, thus resulting in a large number of immunes. Maller and Zhou [10] quoted a study on recidivism times of prisoners released from prisons in Western Australia. The data showed that a significant proportion of prisoners were unlikely to return to prison. The results of the analysis indicate that prison programs and other factors such as age, job status,

and marital status have an effect on recidivism rates. Struthers and Farewell [13] modeled the progression of AIDS in HIV-positive individuals with a cure rate model. They noted that the model that allowed the proportion of immunes provided a better fit to the data than a model which did not account for long term survivors.

The cure rate model has also been used in reliability. Nelson [11] observed the life of insulation on electric motors which were operated at various levels of temperature. He found that the motors lasted almost indefinitely when operated under low temperature, and broke down quickly at higher temperature. Nelson applied the mixture model to capture the immune components in this study. Further research in this area includes articles by Yu *et al.* [16], Farewell [5], Gamel *et al.* [7], Yamaguchi [15], Cancho and Bolfarine [2] and Taylor [14].

Yu *et al.* [16] used the mixture cure rate model for grouped survival data and observed that the estimate of the cure fraction can be quite sensitive to the length of follow up time and the choice of latency distribution. They investigated the effect of various distributions such as the lognormal, loglogisitc, Weibull and generalized gamma, and concluded that the estimate of the cure fraction was robust with the generalized gamma distribution. Yu *et al.* [16] also investigated the identifiability of mixture models, and noted that the overall survival function and the survival function for the latency distribution can become unidentifiable if the follow up time is short. They suggest that a longer follow up time with respect to the median survival time and homogeneity of the observations affect the accuracy of the estimate of the cure fraction.

Yamaguchi [15] proposed an accelerated failure-time regression model with an additional regression model for the cure fraction to study inter-firm job mobility in Japan. He used the generalized gamma to model the latency distribution and the logistic function to model the cure fraction in terms of covariates. This model helps to estimate simultaneously the effect

of covariates on the acceleration (deceleration) of an event as well as the surviving fraction. Farewell [5] also used covariates to model the cure fraction. Chen *et al.* [3] proposed a new Bayesian model for survival data with a surviving fraction. This model has a proportional hazards structure, with the cure rate depending naturally on covariates. One key difference between the the Bayesian approach and the cure rate approach is that the former models the entire population as a proportional hazards model while the later models only the non-cured group with a proportional hazards model. Both the models can be obtained from one another. The authors suggest that the Bayesian model is computationally attractive.

## 1.1 Cure Rate Model

To introduce the cure rate model, we assume that the population consists of two types of patients: susceptibles and immunes. Susceptibles refer to those in the population who are subject to the event of interest such as recurrence of a disease or death. Immunes are those who are not subject to the event of interest and whose survival time is indefinite with respect to the event of interest. These patients survive till the end of the experiment if they do not die from other causes. Let $T$ denote the survival time of an individual. Define an indicator variable $B$, with $B = 0$ when the subject is susceptible, and $B = 1$ when the subject is immune. Let $P(B = 1) = p$, and $P(B = 0) = 1 - p$. Let $F$ denote the cumulative distribution function (cdf) of the overall population and $F_0$ denote the cdf of susceptibles. We assume $F_0$ to be a proper cdf. Then, for a finite $t \geq 0$

$$P(T \leq t|B = 0) = F_0(t) \qquad \text{and} \qquad P(T \leq t|B = 1) = 0.$$

The cdf of the overall population is

$$\begin{aligned} F(t) = P(T \leq t) &= P(T \leq t|B = 0)P(B = 0) + P(T \leq t|B = 1)P(B = 1) \\ &= (1 - p)F_0(t). \end{aligned} \tag{1}$$

Equivalently,

$$F_0(t) = \frac{F(t)}{1-p},$$

which is a rescaled version of $F$. Notice that $F$ is an improper cdf with $F(\infty) < 1$. Let $S(t) = 1 - F(t)$ and $S_0(t) = 1 - F_0(t)$ be the survival functions corresponding to the cdf's $F$ and $F_0$ respectively. Then

$$S(t) = 1 - (1-p)F_0(t) = p + (1-p)S_0(t). \tag{2}$$

The rest of the paper is organized as follows. In section 2, we introduce the Generalized Exponential (GE) distribution and discuss briefly its properties. We formulate the cure rate model based on the GE distribution incorporating the effects of covariates on the probability of being immune. In section 3, we present estimation of the parameters based on the maximum likelihood method. The likelihood equations are solved iteratively using the EM algorithm. In section 4, we present an example illustrating the procedure. Conclusions are presented in section 5. The second order partial derivatives of the likelihood function, which are used to obtain the observed information matrix, are included in the Appendix

## 2 Model Assumptions

In this section, we assume that the distribution of lifetimes for the susceptible population follows a generalized exponential distribution with probability density function given by

$$f(t;\alpha,\lambda) = \alpha\lambda e^{-\lambda t}\left(1 - e^{-\lambda t}\right)^{\alpha-1}; \quad t > 0. \tag{3}$$

Here $\alpha > 0$ and $\lambda > 0$ are the shape and scale parameters respectively. The density function is unimodal and for fixed $\lambda$, it becomes less positively skewed and more negatively skewed as $\alpha$ increases. Note that when $\alpha = 1$, the density function reduces to that of the exponential distribution. The generalized exponential distribution shares many properties similar to

those of the Weibull and Gamma distributions. The hazard function has the same behavior as that of the Gamma distribution, and differs from that of the Weibull. The fact that it has a closed form cdf makes it an attractive alternative to the gamma. See Gupta and Kundu [8] for properties of this distribution and applications. We denote the distribution function and the survival function of the generalized exponential distribution by $F_0(t;\alpha,\lambda)$ and $S_0(t;\alpha,\lambda)$ respectively. We have

$$F_0(t;\alpha,\lambda) = \left(1 - e^{-\lambda t}\right)^{\alpha}. \tag{4}$$

Substituting (4) in (1), we observe that

$$\left(\frac{F(t)}{(1-p)}\right)^{\frac{1}{\alpha}} = 1 - e^{-\lambda t} \Longrightarrow -ln\left[1 - \left(\frac{F(t)}{(1-p)}\right)^{\frac{1}{\alpha}}\right] = \lambda t. \tag{5}$$

If the values of $\alpha$ and $p$ are known, then the plot of $g(\hat{F}(t)) = ln\left[1 - \left(\frac{F(t)}{(1-p)}\right)^{\frac{1}{\alpha}}\right]$ against $t$ should be approximately linear. Here $\hat{F}(t)$ is the Kaplan Meier estimator of $F(t)$. We assume an initial value for $p$ (the Kaplan Meier estimator has been used as an initial guess) and calculate $g(\hat{F}(t))$ using $\alpha = 1$. If the plot of $g(\hat{F}(t))$ versus $t$ is linear, that suggests an exponential distribution is viable, otherwise, we find the value of $\alpha$ for which the plot $g(\hat{F}(t))$ versus $t$ is approximately linear. This procedure will provide initial estimates of $\alpha$ and $\lambda$ that may be used in the iterative methods for determining the MLE's.

We further assume that the probability that an individual is immune depends on a set of covariates $\boldsymbol{z}$. For each individual under study with covariate $\boldsymbol{z}$, we define a binary random variable $\Delta(\boldsymbol{z})$ taking values 0 and 1. It takes the value 0 if the corresponding individual belongs to the immune group and 1 otherwise. We assume that

$$P(\Delta(\boldsymbol{z}) = 0) = p(\boldsymbol{\beta}, \boldsymbol{z}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{z}}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{z}}}, \qquad P(\Delta(\boldsymbol{z}) = 1) = 1 - p(\boldsymbol{\beta}, \boldsymbol{z}) = \frac{1}{1 + e^{\boldsymbol{\beta}'\boldsymbol{z}}}. \tag{6}$$

# 3 Estimation of the Parameters

Our problem is to estimate the unknown parameters, namely $\alpha$, $\lambda$ and $\boldsymbol{\beta}$. We assume that for the $i$-th individual, $t_i$ represents the survival time or the censored time and $\boldsymbol{z_i}$ represents the corresponding vector of covariates. Without loss of generality, we assume that $t_1, \ldots, t_m$ are the actual survival times and $t_{m+1}, \ldots, t_{m+n}$ are the censored times. Based on the above observations the log-likelihood function takes the following form

$$L(\alpha, \lambda, \boldsymbol{\beta}) = L_1(\alpha, \lambda, \boldsymbol{\beta}) + L_2(\alpha, \lambda, \boldsymbol{\beta}),$$

where

$$L_1(\alpha, \lambda, \boldsymbol{\beta}) = \sum_{i=1}^{m} \ln(1 - p(\boldsymbol{\beta}, \boldsymbol{z_i})) + \ln\left(\alpha\lambda \left(1 - e^{-\lambda t_i}\right)^{\alpha-1} e^{-\lambda t_i}\right)$$

and

$$L_2(\alpha, \lambda, \boldsymbol{\beta}) = \sum_{i=m+1}^{m+n} \ln\left(p(\boldsymbol{\beta}, \boldsymbol{z_i}) + (1 - p(\boldsymbol{\beta}, \boldsymbol{z_i}))\left(1 - e^{-\lambda t_i}\right)^{\alpha}\right).$$

The MLE's are obtained by treating this as a missing data problem and using the EM algorithm as follows: for each individual with covariate $\boldsymbol{z}$, the random variable $\Delta(\boldsymbol{z})$ is 1 for the first $m$ individuals, and is unknown for the remaining $n$ individuals. These $n$ observations are treated as missing.

In the 'E' step of the EM algorithm, we compute the pseudo log-likelihood function based on the missing observations. For a censored time $t$, we construct two partially complete 'pseudo observations' of the form $(t, w_1(\boldsymbol{z}, t))$ and $(t, w_2(\boldsymbol{z}, t))$. Specifically, $w_1(\boldsymbol{z}, t)$ and $w_2(\boldsymbol{z}, t)$ denote the conditional probabilities that the individual belongs to the immune or susceptible group, given survival until time $t$. We have

$$w_1(\boldsymbol{z}, t) = P(\Delta(\boldsymbol{z}) = 0 | T > t), \qquad w_2(\boldsymbol{z}, t) = P(\Delta(\boldsymbol{z}) = 1 | T > t).$$

We can write

$$\begin{aligned} w_1(\boldsymbol{z},t) = P(\Delta(\boldsymbol{z})=0|T>t) &= \frac{P(T>t|\Delta(\boldsymbol{z})=0)\times P(\Delta(\boldsymbol{z})=0)}{P(T>t)} \\ &= \frac{p(\boldsymbol{\beta},\boldsymbol{z})}{p(\boldsymbol{\beta},\boldsymbol{z})+(1-p(\boldsymbol{\beta},\boldsymbol{z}))S_0(t)}. \end{aligned}$$

Similarly,

$$\begin{aligned} w_2(\boldsymbol{z},t) = P(\Delta(\boldsymbol{z})=1|T>t) &= \frac{P(T>t|\Delta(\boldsymbol{z})=1)\times P(\Delta(\boldsymbol{z})=1)}{P(T>t)} \\ &= \frac{(1-p(\boldsymbol{\beta},z))S_0(t;\alpha,\lambda)}{p(\boldsymbol{\beta},\boldsymbol{z})+(1-p(\boldsymbol{\beta},\boldsymbol{z}))S_0(t;\alpha,\lambda)}. \end{aligned}$$

The 'pseudo log-likelihood', $L_{pseudo}(\alpha,\lambda,\boldsymbol{\beta})$ based on the missing observations is;

$$\begin{aligned} L_{pseudo}(\alpha,\lambda,\boldsymbol{\beta}) &= L_1(\alpha,\lambda,\boldsymbol{\beta}) \\ &+ \sum_{i=m+1}^{m+n}\{w_1(\boldsymbol{z_i})\ln p(\boldsymbol{\beta},\boldsymbol{z_i}) + w_2(\boldsymbol{z_i})\ln[(1-p(\boldsymbol{\beta},\boldsymbol{z_i}))S_0(t_i;\alpha,\lambda)]\} \\ &= g_1(\boldsymbol{\beta}) + g_2(\alpha,\lambda), \quad \text{(say)} \end{aligned} \tag{7}$$

where

$$g_1(\boldsymbol{\beta}) = \sum_{i=1}^{m}\ln(1-p(\boldsymbol{\beta},\boldsymbol{z_i})) + \sum_{i=m+1}^{m+n} w_1(\boldsymbol{z_i})\ln p(\boldsymbol{\beta},\boldsymbol{z_i}) + \sum_{i=m+1}^{m+n} w_2(\boldsymbol{z_i})\ln(1-p(\boldsymbol{\beta},\boldsymbol{z_i}))$$

and

$$g_2(\alpha,\lambda) = \sum_{i=1}^{m}\ln f(t_i;\alpha,\lambda) + \sum_{i=m+1}^{m+n} w_2(\boldsymbol{z_i})\ln S_0(t_i;\alpha,\lambda).$$

The 'M' step of the EM algorithm involves maximizing $L_{pseudo}(\alpha,\lambda,\boldsymbol{\beta})$ with respect to the unknown parameters for fixed $w_1(\boldsymbol{z})$ and $w_2(\boldsymbol{z})$. Since the pseudo log-likelihood function $L_{pseudo}(\alpha,\lambda,\boldsymbol{\beta})$ can be written as (7), therefore, if $\alpha^{(k)}$, $\lambda^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ are estimates of $\alpha$, $\lambda$ and $\boldsymbol{\beta}$ at the $k-th$ iterate then $\boldsymbol{\beta}^{(k+1)}$ can be obtained by maximizing $g_1(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and $\alpha^{(k+1)}$, $\lambda^{(k+1)}$ can be obtained by maximizing $g_2(\alpha,\lambda)$ with respect to $\alpha$ and $\lambda$ respectively for fixed $w_1(\boldsymbol{z})$ and $w_2(\boldsymbol{z})$. Note that for the $(k+1)-th$ step, $w_1(\boldsymbol{z})$ and $w_2(\boldsymbol{z})$ depend on $\alpha^{(k)}$, $\lambda^{(k)}$ and $\beta^{(k)}$ and they are as follows;

$$w_1(\boldsymbol{z}) = \frac{p(\beta^{(k)},\boldsymbol{z})}{p(\beta^{(k)},\boldsymbol{z})+(1-p(\beta^{(k)},\boldsymbol{z}))S_0(t;\alpha^{(k)},\lambda^{(k)})},$$

$$w_2(\boldsymbol{z}) = \frac{(1-p(\beta^{(k)},\boldsymbol{z}))S_0(t;\alpha^{(k)},\lambda^{(k)})}{p(\beta^{(k)},\boldsymbol{z})+(1-p(\beta^{(k)},\boldsymbol{z}))S_0(t;\alpha^{(k)},\lambda^{(k)})}.$$

Now we can maximize $g_1(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and $g_2(\alpha,\lambda)$ with respect to $\alpha$ and $\lambda$ separately.

REMARK 1: If $\beta$ is a scalar then,

$$\begin{aligned} g_1(\beta) &= \sum_{i=1}^{m} \ln \frac{1}{1+e^{\beta z_i}} + \sum_{i=m+1}^{m+n} w_2(z_i) \ln \frac{1}{1+e^{\beta z_i}} + \sum_{i=m+1}^{m+n} w_1(z_i) \ln \frac{e^{\beta z_i}}{1+e^{\beta z_i}} \\ &= -\sum_{i=1}^{m+n} \ln(1+e^{\beta z_i}) + \beta \sum_{i=m+1}^{m+n} z_i w_1(z_i). \end{aligned}$$

Therefore, the maximization can be obtained by differentiating $g_1(\beta)$ with respect to $\beta$ and equating to 0, *i.e.* by solving the following nonlinear equation

$$g_1'(\beta) = -\sum_{i=1}^{m+n} \frac{z_i e^{\beta z_i}}{1+e^{\beta z_i}} + \sum_{i=m+1}^{m+n} z_i w_1(z_i) = 0.$$

Since

$$g_1''(\beta) = -\sum_{i=1}^{m+n} \frac{z_i^2 e^{\beta z_i}}{(1+e^{\beta z_i})^2} < 0,$$

therefore $g(\beta)$ is a concave function. ∎

REMARK 2: If $\boldsymbol{\beta}$ is not a scalar, it can be shown that the matrix $\dfrac{\partial^2 g_1(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}$ is negative definite under mild restrictions on $\boldsymbol{z}$'s. ∎

For the GE model, we have

$$\begin{aligned} g_2(\alpha,\lambda) &= m\ln\alpha + m\ln\lambda - \lambda\sum_{i=1}^{m} t_i + (\alpha-1)\sum_{i=1}^{m}\ln(1-e^{-\lambda t_i}) \\ &+ \sum_{i=m+1}^{n} w_2(z_i)\ln(1-(1-e^{-\lambda t_i})^{\alpha}). \end{aligned}$$

The method proposed by Song *et al.* [12] can be used to maximize $g_2(\alpha,\lambda)$. Let us write

$$g_2(\alpha,\lambda) = h_1(\alpha,\lambda) + h_2(\alpha,\lambda),$$

where

$$h_1(\alpha,\lambda) = m\ln\alpha + m\ln\lambda - \lambda\sum_{i=1}^{m} t_i + (\alpha-1)\sum_{i=1}^{m}\ln(1-e^{-\lambda t_i})$$

and

$$h_2(\alpha,\lambda)=\sum_{m+1}^{n} w_2(z_i)\ln(1-(1-e^{-\lambda t_i})^{\alpha}).$$

Since we need to solve

$$g_2'(\alpha,\lambda)=h_1'(\alpha,\lambda)+h_2'(\alpha,\lambda)=0$$

or

$$h_1'(\alpha,\lambda)=-h_2'(\alpha,\lambda),$$

we use the following procedure. First solve

$$h_1'(\alpha,\lambda)=0.$$

using the following non-linear equation (fixed point type) iteratively

$$\lambda=\left(\frac{1}{m}\sum_{i=1}^{m}\frac{t_i e^{-\lambda t_i}}{(1-e^{-\lambda t_i})}\left(1+\frac{m}{\sum_{i=1}^{m}\ln(1-e^{-\lambda t_i})}\right)+\frac{1}{m}\sum_{i=1}^{m}t_i\right)^{-1}.$$

If $\lambda^{(0)}$ is the solution then obtain

$$\alpha^{(0)}=-\frac{m}{\sum_{i=1}^{m}\ln(1-e^{-\lambda^{(0)}t_i})}.$$

Now obtain $\alpha^{(1)}$ and $\lambda^{(1)}$ as the solution of the following

$$h_1'(\alpha,\lambda)=-h_2'(\alpha^{(0)},\lambda^{(0)}).$$

Once we obtain $\alpha^{(1)}$ and $\lambda^{(1)}$ then $\alpha^{(2)}$ and $\lambda^{(2)}$ as the solution of the following

$$h_1'(\alpha,\lambda)=-h_2'(\alpha^{(1)},\lambda^{(1)}).$$

It should be continued until it converges. Note that the solution $(\tilde{\alpha},\tilde{\lambda})$ of the following equation, for any arbitrary $c_1$ and $c_2$

$$h_1'(\alpha,\lambda)=(c_1,c_2),$$

can be obtained as follows. First solve the non-linear equation iteratively

$$\lambda=\left[\frac{c_2}{m}+\frac{1}{m}\sum_{i=1}^{m}t_i+\left(1-\frac{m}{c_1-\sum_{i=1}^{m}\ln(1-e^{-\lambda t_i})}\right)\times\left(\frac{1}{m}\sum_{i=1}^{m}\frac{t_i e^{-\lambda t_i}}{1-e^{-\lambda t_i}}\right)\right]^{-1},$$

to obtain $\tilde{\lambda}$ and then obtain,

$$\tilde{\alpha} = \left[ \frac{c_1 - \sum_{i=1}^{m} \ln(1 - e^{-\lambda t_i})}{m} \right]^{-1}.$$

# 4 Simulation and Data Analysis

## 4.1 Simulation Results

In this section, we present results of a small simulation study to see how the proposed EM algorithm works for different sample sizes and parameter values. For simulation purposes we have used the cure rate model without any covariates. We have taken different $\alpha$ and $n$ values. We report the average values of the estimates and the mean squared errors over 1000 replications in Table 1.

Table 1: Average estimates and mean squared errors

| $n$ | $\alpha = 2.0$ | $\lambda = 1$ | $p = 0.1$ | $n$ | $\alpha = 1.0$ | $\lambda = 1$ | $p = 0.1$ |
|---|---|---|---|---|---|---|---|
| 25 | 2.1363 | 1.2671 | 0.1213 | 25 | 1.2384 | 1.4316 | 0.1165 |
| | (0.6762) | (0.1728) | 0.0073) | | (0.1611) | (0.2881) | 0.0074) |
| 50 | 1.9424 | 1.2004 | 0.1137 | 50 | 1.1693 | 1.3808 | 0.1145 |
| | (0.3789) | (0.1005) | (0.0057) | | (0.0613) | ( 0.2165) | (0.0053) |
| 75 | 1.8808 | 1.1850 | 0.1132 | 75 | 1.1383 | 1.3721 | 0.1141 |
| | ( 0.2491) | ( 0.0756) | (0.0034) | | (0.0360) | (0.1938) | ( 0.0036) |

From the simulation study, we observe that the EM algorithm converges quite fast and the average biases and mean squared errors converge to zero as the sample size increases. It is interesting to observe that the true value of $\alpha$ affects the estimation of $\lambda$ but does not have much of an effect on the estimation of $p$.

## 4.2 Data Analysis

In this section, we analyze a subset of data from the University of Massachusetts Aids Research Unit Impact Study. The data is available from the website of statistical data at the University of Massachusetts/ Amherst ( http://www-unix.oit.umass.edu/ statdata).

The data includes results from randomized trials of two different residential treatment programs aimed at reducing drug abuse and consequently high-risk behavior. There were two treatment program sites referred to as site 1 and site 2. At site 1, participants were randomized to 3 and 6 month groups that included health education and relapse prevention programs. Site 2 participants were assigned to a structured life-style environment and randomized to 6 or 12 month programs. The "survival" time here refers to the number of days from admission to the time the participant returned to drug use (self-reported). We created 4 groups based on the Site (S) X Length (L) assignment as follows:

| S → <br> L ↓ | One | Two |
|---|---|---|
| Short | 1 | 2 |
| Long | 3 | 4 |

There is evidence among researchers to suggest that a proportion of participants in drug and alcohol treatment programs are "cured", i.e. will never abuse drugs or alcohol again. The probability of being "cured" is usually affected by factors such as age, environment, and employment status. It is therefore reasonable to represent the data using the cure-rate model. We analyzed the data consisting of 575 observations. We used three covariates in the analysis: Age at enrollment in years (AGE), IV Drug use history at admission (IV), and Number of drug treatments (NDRUG). The IV drug use history was treated as a binary covariate: the value 1 assigned to individuals who had a recent history of IV drug use, and the value 0 assigned to individuals with no recent history of IV drug use. The number of

prior drug treatments took values from 0 (for no prior treatments) to 40 and was treated as a categorical variable in the analysis.

In our analysis of the model, we assumed that the covariates affect only the probability of being cured. We do not consider the more general model wherein the covariates may affect both the probability of being cured as well as the survival distribution of the susceptible population. Inference for the general model will be significantly more complex and issues of identifiability will need to be resolved.

Table 2 provides the estimates of the parameters $p, \alpha$ and $\lambda$ of the generalized exponential model for four groups without any covariate information. The point estimate value and the corresponding margin of error (for a 95 % interval) along with the log-likelihood value are reported. The intervals for $\alpha$ for Groups 2 and 4 indicate that the simple exponential model may be adequate.

Table 2: Parameter Estimates without Covariates

| Group | $n$ | $m$ | $\widehat{\alpha}$ | $\widehat{\lambda}$ | $\widehat{p}$ | LL |
|---|---|---|---|---|---|---|
| I | 198 | 167 | 1.6369 (0.5473) | 0.4721 (0.1577) | 0.1555 (0.0522) | -208.096 |
| II | 91 | 72 | 1.5658 (0.6897) | 0.4018 (0.1777) | 0.2051 (0.0872) | -97.928 |
| III | 202 | 159 | 1.4986 (0.4337) | 0.3274 (0.1054) | 0.2056 (0.1284) | -205.723 |
| IV | 84 | 66 | 1.2874 (0.4862) | 0.2731 (0.1054) | 0.2018 (0.1284) | -75.981 |

We also provide a test for goodness of fit for Group 1 using the K-M estimate of the survival function. Figure 1 shows the transformed K-M estimate plotted against time for $\alpha = 1$ (the exponential model), and $\alpha = 1.637$. Clearly, the generalized exponential model provides an excellent fit to the data. Similar results are obtained for Group 3.

Figure 2 shows an overlay plot of the K-M estimator of the survival function for Group

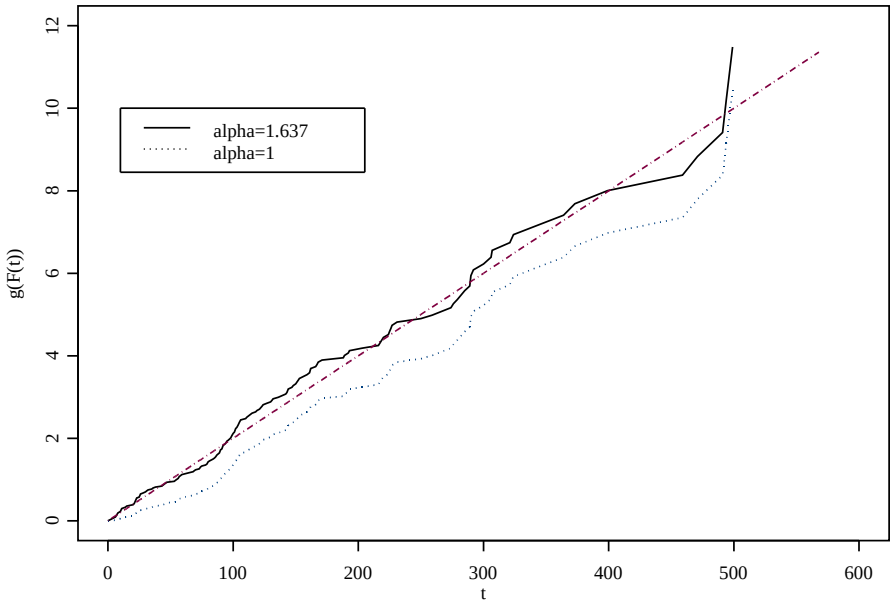


Figure 1: Plot of $g(\hat{F}(t))$ versus $t$ for 2 choices of $\alpha$

1 and the estimated survival function based on the GE model. This graph provides further support that the GE model does indeed provide an excellent fit to the data and captures the immune proportion.

## 4.3 Covariate Model

Next, we fit the model using 3 covariates: Age, IV, NDRUG. Table 3 provides the estimates of the regression coefficients. The point estimate and the corresponding margin of error (for a 95% interval) along with the loglikelihood values are reported. All three covariates are highly significant. The log-likelihood values may be used to assess the improved fit of the model using the standard chi-square criterion. The significance of all three covariates indicate that the probability of being "cured" depends on the persons recent drug history and age. Patients who have been through several previous rehab programs seem to have

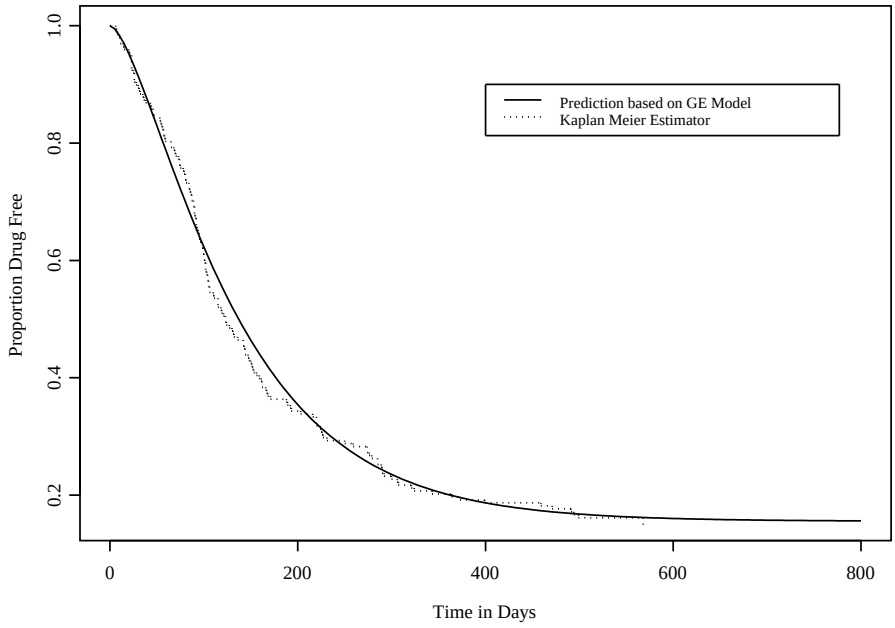


Figure 2: Kaplan Meier Estimator and Predicted Model for Group 1

a higher chance of relapse. These findings are consistent with experts in the area of drug prevention and addiction.

For the binary covariate IV, all the parameter estimates are negative. For example, in Group 1, the odds of being cured decrease by 50% for recent IV drug users. Similarly, the odds of being cured decrease by 85% for a unit increase in the number of previous drug treatments.

Figure 3 provides the predicted survival functions for Group 1 and illustrates the effects of recent IV use. Clearly, for individuals who have a recent history of IV use, the survival curve is steeper and the cured proportion significantly lower. Figure 4 provides the predicted survival functions for the 4 groups for a fixed set of covariate values. It is clear from Figure 4 that in all the four cases the cured proportions are approximately same.

Table 3: Parameter Estimates for the Model with Covariates

| Group | $\widehat{\beta}_0$ | $\widehat{\beta}_{AGE}$ | $\widehat{\beta}_{IV}$ | $\widehat{\beta}_{NDRUG}$ | LL |
|---|---|---|---|---|---|
| I | -2.4250 | 2.5501 | -0.6251 | -4.9498 | -198.442 |
| | (1.239) | (1.914) | (0.441) | (2.225) | |
| II | -0.6895 | -.4510 | -1.0991 | -1.4891 | -93.262 |
| | (0.2823) | (0.3359) | (0.8230) | (1.3442) | |
| III | -3.5500 | 4.3489 | -0.4511 | -3.2112 | -197.434 |
| | (1.0599) | (1.6039) | (0.3897) | (1.5589) | |
| IV | -2.5486 | 2.7512 | -0.9976 | -4.7001 | -72.050 |
| | (1.5512) | (2.3462) | (0.8751) | (3.3670) | |

# 5 Conclusions

In this article, we investigated the performance of the cure rate model based on the generalized exponential distribution. The probability of being immune was modeled as a function of covariates using the logistic function. The parameter estimates were obtained via the EM algorithm. We illustrated the performance of the model by using data from a study on drug treatment programs. The results indicate the GE model provides an excellent fit to the data. We were also able to test the effects of different covariates on the immune probability.

The model proposed in this paper assumed that the covariates affect only the probability of being cured. A more general model may be proposed that also includes the effects of covariates on the failure time distribution. Covariates may be classified into three groups: the first group containing covariates that affect only the survival distribution, the second group containing covariates that affect only the immune probability, and the third group containing covariates that affect both. For covariates that appear in both $p$ and $S_0(t)$, there are potential issues of interpretation and identifiability that would need to be addressed. In addition, the estimation of the parameters and associated inference will be more complex. Further investigation is needed to address these issues in the cure rate framework.

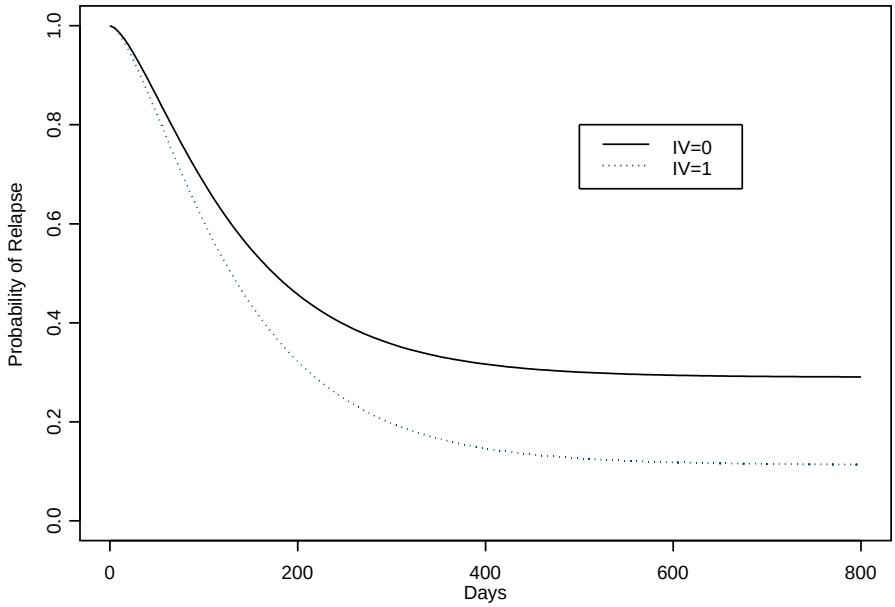


Figure 3: Plot of Predicted Survival Functions: Age=30, IV=1, NDRUG=0

## Acknowledgements

The authors would like to thank the associate editor and two referees for their valuable comments, which has improved the earlier version of the manuscript. Part of this work is supported by a grant from the Department of Science and Technology, Government of India.

# References

[1] Boag, J. W. (1949), "Maximum likelihood estimates of the proportion of patients cured by cancer therapy", *Journal of the Royal Statistical Society, Series B*, vol. 11, 15–53.

[2] Cancho, V.G. and Bolfarine, H. (2001), "Modeling the presence of immunes by using the exponentiated-Weibull model", *Journal of Applied Statistics*, vol. 28, 659 - 671.

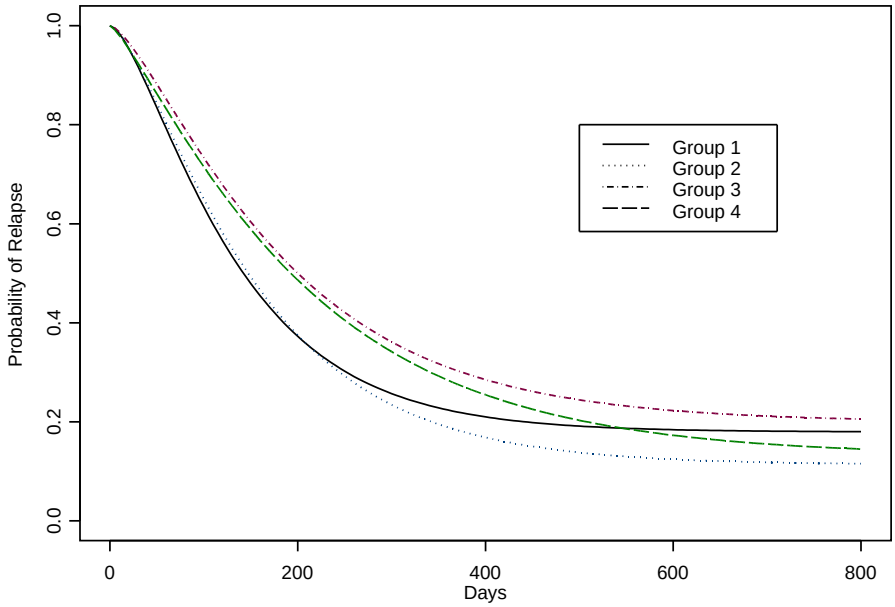


Figure 4: Plot of Predicted Survival Functions: Age=30, IV=1, NDRUG=0

[3] Chen Ming-Hui, Ibrahim J., Sinha D. (1999), "A new Bayesian model for survival data with a surviving fraction", *Journal of the American Statistical Association*, vol. 94, 909 - 918.

[4] Dunsmuir, W. Tweedie, R. Flack, L. and Mengersen, K. (1989), "Modeling of transitions between employment states for young Australians" *Australian Journal of Statistics*, vol. 31, A,165–196.

[5] Farewell, V. T. (1982), " The use of mixture models for the analysis of survival data with long-term survivors", *Biometrics*, vol. 38, 1041–1046.

[6] Freireich, E. J., Gehan, E., Frei, E., Schroeder, L. R., Wolman, I. J., Anbari, R., Burgert, E. O., Mills, S. D., Pinkel, D., Selawry, O. S., Moon, J. H., Gendel, B. R., Spurr, C. L., Storrs, R., Haurani, F., Hoogstraten, B. and Lee, S. (1963), "The effect

of 6-Mercaptopurine on the duration of steroid-induced remissions in acute leukemia; a model for evaluation of other potentially useful therapy", *Blood*, vol. 21, 699–716.

[7] Gamel, J. W., Mclean, I. W, and Rosenberg, S. H. (1999), "Proportion cured and mean log-survival time as functions of tumor size", *Statistics in Medicine*, vol. 9, 999-1006.

[8] Gupta, R. D. and Kundu, D. (1999), "Generalized exponential distributions", *Australian and New Zealand Journal of Statistics*, vol. 41, 173–188.

[9] Louis, T.A., "Finding the observed information matrix when using the EM algorithm", *Journal of the Royal Statistical Society*, B, vol. 44, 226 - 233.

[10] Maller, R, and Zhou X. (1996), *Survival analysis with long-term survivors*, John Wiley & Sons, Inc., New York.

[11] Nelson, W. (1982) *Applied life data analysis*, John Wiley & Sons, New York.

[12] Song, P.X., Fan, Y., Kalbfleisch, J.D. (2005), "Maximization by parts in likelihood inference (with discussions)", *Journal of the American Statistical Association*, vol. 100, 11451167.

[13] Struthers, C. A. and Farewell, V. T. (1989), "A mixture model for time to AIDS data with left truncation and an uncertain origin", *Biometrika*, vol. 76, 814-817.

[14] Taylor, J. M. G. (1995), "Semiparametric estimation in failure time mixture models", *Biometrics*, vol. 51, 899-907.

[15] Yamaguchi, K. (1992), "Accelerated failure-time regression model with a regression model for the surviving fraction: an application to the analysis of 'permanent employment' in Japan", *Journal of the American Statistical Association*, vol. 87, 284-292.

[16] Yu, B., Tiwari, R. C. and Cronin, K. Z. (2004), "Cure fraction estimation from the mixture cure models for grouped survival times", *Statistics in Medicine*, vol. 23, 1733-1747.

# Appendix

### Observed Fisher Information Matrix

In this case we can use the idea of Louis [9], to compute the observed Fisher Information matrix $\hat{\mathbf{I}}$. Using the same notation of Louis [9], it can be observed that $\hat{\mathbf{I}}$ takes the form

$$\hat{\mathbf{I}} = \hat{\mathbf{B}} - \hat{\mathbf{S}}\hat{\mathbf{S}}^T. \tag{8}$$

Here $\hat{\mathbf{B}}$ is the $2+k+1 \times 2+k+1$ negative of the second derivative matrix and $\hat{\mathbf{S}}$ is the $2+k+1$ gradient vector and $k =$ the number of covariates used in the model. We decompose $\hat{\mathbf{B}}$ and $\hat{\mathbf{S}}$ as follows;

$$\hat{\mathbf{B}} = \begin{bmatrix} \hat{\mathbf{B}}_{11} & \hat{\mathbf{B}}_{12} \\ \hat{\mathbf{B}}_{21} & \hat{\mathbf{B}}_{22} \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{S}} = \begin{bmatrix} \hat{\mathbf{S}}_1 \\ \hat{\mathbf{S}}_2 \end{bmatrix},$$

where $\hat{\mathbf{B}}_{11}$ and $\hat{\mathbf{B}}_{22}$ are $2 \times 2$ and $k+1 \times k+1$ matrices, similarly, $\hat{\mathbf{S}}_1$ and $\hat{\mathbf{S}}_2$ are $2 \times 1$ and $k+1 \times 1$ vectors respectively. It can be easily observed that $\hat{\mathbf{B}}_{12} = \hat{\mathbf{B}}_{21}^T = \mathbf{0}$. We will be using the following notation. The $(i,j)$-th elements of the matrices of $\hat{\mathbf{B}}_{11}$ and $\hat{\mathbf{B}}_{22}$ will be denoted by $((b_{11}(i,j)))$ and $((b_{22}(i,j)))$ respectively. Similarly, the $i$-th elements of $\hat{\mathbf{S}}_1$ and $\hat{\mathbf{S}}_2$ will be denoted by $\hat{S}_1(i)$ and $\hat{S}_2(i)$ respectively. Moreover, for the $i$-th individual the covariate is denoted by $\boldsymbol{z_i} = \{1, z_{1i}, \ldots, z_{ki}\}$ and

$$\hat{p}(i) = \frac{e^{\hat{\boldsymbol{\beta}}'\boldsymbol{z_i}}}{1+e^{\hat{\boldsymbol{\beta}}'\boldsymbol{z_i}}}, \quad \hat{w}_1(i) = \frac{\hat{p}(i)}{\hat{p}(i)+(1-\hat{p}(i))S(t;\hat{\alpha},\hat{\lambda})} \quad \hat{w}_2(i) = 1-\hat{w}_1(i),$$

where $\hat{\boldsymbol{\beta}}$, $\hat{\alpha}$ and $\hat{\lambda}$ are the respective MLEs. Moreover we denote $a_i = \left(1-e^{\hat{\lambda}t_i}\right)$ and $b_i = \left(1-e^{\hat{\lambda}t_i}\right)^{\hat{\alpha}}$

With the above notation, it can be observed after some lengthy calculations that

$$
\begin{aligned}
b_{11}(1,1) &= \frac{m}{\hat{\alpha}^2} + \sum_{i=m+1}^{m+n} \frac{\hat{w}_2(i) b_i (\ln a_i)^2}{(1-b_i)^2} \\
b_{11}(2,2) &= \frac{m}{\hat{\lambda}^2} + (\hat{\alpha}-1) \sum_{i=1}^{m} \frac{t_i^2 e^{-\hat{\lambda} t_i}}{a_i^2} + \hat{\alpha} \sum_{i=m+1}^{m+n} \frac{\hat{w}_2(i) t_i^2 e^{-\hat{\lambda} t_i} b_i}{a_i^2 (1-b_i)^2} \{\hat{\alpha} e^{-\hat{\lambda} t_i} - 1 + b_i\} \\
b_{11}(1,2) &= b_{11}(2,1) = -\sum_{i=1}^{m} \frac{1}{a_i} t_i e^{-\hat{\lambda} t_i} - \sum_{i=m+1}^{m+n} \frac{\hat{w}_2(i) t_i e^{-\hat{\lambda} t_i} b_i}{a_i (1-b_i)^2} \{a_i + \hat{\alpha} \ln a_i - a_i b_i\} \\
b_{22}(1,1) &= \sum_{i=1}^{m+n} \hat{p}(i)(1-\hat{p}(i)) \\
b_{22}(j,j) &= \sum_{i=1}^{m+n} z_{ji}^2 \, \hat{p}(i)(1-\hat{p}(i)) \quad \text{for} \quad j = 2, \ldots k+1 \\
b_{22}(1,j) &= \sum_{i=1}^{m+n} z_{ji} \hat{p}(i)(1-\hat{p}(i)) \quad \text{for} \quad j = 2, \ldots, k+1 \\
b_{22}(l,j) &= \sum_{i=1}^{m+n} z_{ji} z_{li} \, \hat{p}(i)(1-\hat{p}(i)) \quad \text{for} \quad l, j = 2, \ldots k+1 \\
\hat{S}_1(1) &= \frac{m}{\hat{\alpha}} + \sum_{i=1}^{m} \ln a_i - \sum_{i=m+1}^{m+n} \hat{w}_2(i) \frac{b_i \ln a_i}{1-b_i} \\
\hat{S}_1(2) &= \frac{m}{\hat{\lambda}} + (\hat{\alpha}-1) \sum_{i=1}^{m} \frac{1}{a_i} t_i e^{-\hat{\lambda} t_i} - \sum_{i=1}^{m} t_i - \sum_{i=m+1}^{m+n} \hat{w}_2(i) \frac{\hat{\alpha} b_i t_i e^{-\hat{\lambda} t_i}}{a_i (1-b_i)} \\
\hat{S}_2(1) &= -\sum_{i=1}^{m+n} \hat{p}(i) + \sum_{i=m+1}^{m+n} \hat{w}_1(i) \\
\hat{S}_2(j) &= -\sum_{i=1}^{m+n} z_{ji} \, \hat{p}(i) + \sum_{i=m+1}^{m+n} z_{ji} \, \hat{w}_1(i) \quad \text{for} \quad j = 2, \ldots, k.
\end{aligned}
$$