CURE RATE MODELS – A PARTIAL REVIEW WITH AN APPLICATION TO RECURRENT EVENT OR COUNT DATA

Sumathi K[1] , Aruna Rao K[2]

1. A.B.Shetty Memorial Institute of Dental Sciences
   Nitte University, Deralakatte, Mangalore, Karnataka, India.
2. Mangalore University, Karnataka, India.

Abstract :

Cure rate models are survival models consisting of a cured fraction and an uncured fraction. These models are being widely used in analyzing data from cancer clinical trials. A model to estimate the cure fraction was first developed by Boag in 1949 and later developed by Berkson and Gage in 1952 . It was called the mixture model. It is also known as the standard cure rate model. Yakovlev et.al. in 1993 developed an alternative to the mixture model. This model is known as the bounded cumulative hazard (BCH) model. It was developed by considering the number of metastasis competent tumor cells which were left active even after the initial treatment for a cancer patient. The model could overcome some of the drawbacks of the standard cure rate model. Parametric and semi-parametric versions of the two models have been extensively studied. The bivariate extension of the univariate cure rate models include the joint modeling of times to relapse of the disease at two different organs, times to relapse of disease and death, times to occurrence of primary and secondary complications of a disease and joint modeling of time-to-event data and longitudinal data. These extensions involve the use of copulas and frailties. Although some results and applications have been reported, there is a necessity of further work some of which are identified. A cure rate model for count data is proposed and its application has been illustrated in estimating the proportion of people who do not have any dental problems at a given point of time.

Key words : cure rate models, mixture model, bounded cumulative hazard (BCH) model, copulas, count data.

1. INTRODUCTION :

Any clinical trial consists of heterogeneous population of patients which eventually on treatment divides into two groups. One group consists of those patients who respond favourably to the treatment and subsequently become immune or insusceptible to the disease and are said to be cured. The other group consists of those patients who do not respond to the treatment and remain uncured. The main interest of the investigator conducting the clinical trial is in determining the proportion of patients cured and studying the causes for the failure of the treatment in the uncured group of patients. This proportion became an important and a very useful measure in obtaining the trends in the survival of patients suffering from cancer. The widely used model in survival analysis is the Cox (1972) proportional hazards model. This model is based on the assumption that every individual in the population under study is susceptible to the adverse event of interest. This assumption cannot be used in recent clinical trials since a large group of patients are cured of the disease under study, especially cancer, after a sufficient follow up. Thus the mixture model for cure rate is widely used.

Cure models are survival models basically developed to estimate the proportion of patients cured in a clinical trial. These models estimate the cured proportion and also the probability of survival of the uncured patients up to a given point of time. The model developed by Boag (1949) was to estimate the proportion of patients cured among those who were receiving treatment for cancer of mouth and throat, cervix, uteri and breast. This model is called the mixture model since it can estimate the proportion of patients cured and the survival function of the uncured patients. Boag modeled the survival function of the uncured group as a product of the survival functions of a log-normal distribution and some background distribution for the normal population. This model was further developed by Berkson and Gage in 1952 and later studied extensively by several authors. It was observed that the model could be either parametric or non-parametric. The mixture model is said to be a parametric mixture cure model when standard probability distributions such as exponential, Weibull, Gompertz and generalized F are used. The mixture model used without any standard probability distribution is called a non-parametric mixture cure model.

The literature on mixture model is also found in the work by Mendenhall and Hader (1958), Haybittle (1965), Miller (1981), Farewell (1982, 1986), Goldman (1984), Greenhouse and Wolfe (1984), Gray and Tsiatis (1989), Gamel, McLean and Rosenberg (1990), Gordon (1990), Kuk and Chen (1992), Laska and Meisner (1992), Maller and Zhou (1992, 1994, 1995, 1996), Sposto, Sather and Baker (1992), Yamaguchi (1992), Taylor (1995), Peng, Dear and Denham(1998), Angelis, et.al. (1999), Peng and Dear (2000), Sy and Taylor (2000), Betensky and Schoenfeld (2001) and many more.

The mixture model is also known as the standard cure rate model. It was observed that this model had a few limitations especially when a set of covariates were present. Thus an alternative to the standard cure rate model was developed by Yakovlev et.al. (1993). This model is called the bounded cumulative hazard (BCH) model. A brief literature on this model is found in Chapter 5 of the book 'Bayesian Survival Analysis' by Ibrahim, Chen and Sinha (2001).

There exists a vast literature on cure rate models in univariate set-up. The extension of these models to multivariate cases has also been discussed. The applications of cure models include the study of onset of secondary complications of a disease in two or more organs of a patient and the study of onset of several other diseases in an individual who is already suffering from one disease. The cure models are also becoming popular in jointly modeling the overall risk of a disease and the distribution of the age-at-onset of the disease for the diseased individuals.

The objective of this paper is to provide a partial review of cure rate models and to introduce a cure model for recurrent event or count data so as to enable medical practitioners and researchers who lack technical knowledge in using the available softwares to estimate the cure fraction. The paper accounts for the various types of work available on cure rate models which could provide a new researcher the potential areas of research. The rest of the article is as organized as follows. Section 2 discusses three types of univariate cure rate models viz., the mixture model, the BCH model and the model based on Box-Cox transformation developed by Yin and Ibrahim (2005). An extension of

a univariate to a bivariate set-up is possible through copulas. Therefore section 3 provides a brief discussion on copulas. An overview of multivariate cure rate models is given in section 4. Section 5 discusses a cure rate model for recurrent event data with an example from a dental health set up. Conclusions are given in section 6. Some areas of future research are mentioned in section 7.

2. UNIVARIATE CURE RATE MODELS:

2.1 MIXTURE MODEL:

Consider a group of patients entering a clinical trial. Let ϕ be the proportion of patients cured on treatment and 1-ϕ be the proportion uncured. The survivor function *S(t)* for the entire population of patients entering the clinical trial is given by the model

$$S(t) = \phi + (1-\phi)S_u(t) \qquad (2.1.1)$$

where $S_u(t)$ is the survivor function for the uncured group. The commonly used distributions for $S_u(t)$ are the exponential and the Weibull. The model (2.1.1) is known as the mixture cure rate model or the standard cure rate model. The probability density function corresponding to (2.1.1) is

$$f(t) = (1- \phi)f_u(t) \qquad (2.1.2)$$

and the hazard function is

$$h(t) = \frac{(1-\phi) f_u(t)}{\phi + (1-\phi) f_u(t)} \qquad (2.1.3)$$

The density and the survival functions of the cured patients will be equal to 0 and 1 respectively. The mixture model will be parametric or non-parametric depending on whether $f_u(t)$ is specified or not.

Suppose there are *n* patients entering a study. Let t_i , i = 1,2, …., n be the observed survival time for the ith patient. Let y_i be a censoring indicator defined such that

$$y_i = \begin{cases} 1 \text{ if } t_i \text{ is censored} \\ 0 \text{ otherwise} \end{cases}$$

Then the likelihood function is given by

$$L = \prod_{i=1}^{n} \{(1-\phi) f_u(t_i)\}^{y_i} \{\phi + (1-\phi)S_u(t_i)\}^{1-y_i} \qquad (2.1.4)$$

Suppose we have a covariate vector $X_{k\times 1}$ consisting of variables that are likely to influence the health status of the patient, then the dependence of the cure rate ϕ on X can be modeled using a logistic function $\phi = \frac{1}{1+\exp(X^1\beta)}$ where $\beta_{k\times 1}$ is a vector of regression coefficients. The maximum likelihood (ML) estimates can be obtained by maximizing the log-likelihood equation. Newton-Raphson method of iteration can be used if the ML equations do not give a solution in closed form.

Suppose $f_u(t)$ is not specified. Then we have a non-parametric mixture cure model. Maller and Zhou (1992) and Sposto, Sather and Baker (1992) have suggested a non-parametric estimator for the cure rate ϕ as the maximum observed value of the Kaplan-Meir (1958)(KM) estimator given by

$$\hat{S}(t) = \prod_{k:t_{(k)}<t}\left(1-\frac{d_k}{r_k}\right) \qquad (2.1.5)$$

This estimator has been obtained by considering $t_{(1)}<t_{(2)}<t_{(3)}<\ldots\ldots\ldots<t_{(m)}$ the distinct ordered failure times in a group of **n** patients**.** d_k represents the number of failures at time t_k and r_k *i*s the number of patients at risk just before t_k. The KM estimator is also known as the product-limit (PL) estimator.

When a set covariates $X_{k\times 1}$ is present in a non-parametric set-up, the Cox proportional hazards (PH) model

$$h_u(t|\,X) = h_{u0}(t)\exp(X^1\beta) \qquad (2.1.6)$$

is used. Here $h_u(t|X)$ is the hazard function for the uncured patients and $h_{u0}(t)$ is the baseline hazard which can be an arbitrary unspecified hazard function independent of X. The estimation technique based on expectation-maximization (EM) algorithm has been discussed by Peng and Dear (2000).

2.2 BOUNDED CUMULATIVE HAZARD (BCH) MODEL:

The second type of cure model is the BCH model developed by Yakovlev et.al. (1993) as an alternative to the standard cure rate model (2.1.1). The BCH model was developed by considering the patients suffering from cancer entering a clinical trial and was based on the assumption that after treatment a patient is left with *N* cancer cells

capable of metastasizing. The process in which the cancer cells move to various parts of the human body, grow rapidly and replace the normal tissue is known as metastasis. *N* is assumed to follow a Poisson distribution with mean θ. If t_k, $k = 1,2,\ldots,N$, denotes the time for the k^{th} metastatic tumor cell to produce detectable cancer, then the time to relapse of cancer in the patient is given by

$T = minimum\ \{t_1, t_2, \ldots, t_N\}$.

The variable *T* is observable. Conditional on *N*, it is assumed that t_k's are independent and identically distributed (i.i.d.) with cumulative distribution function *F(t),* survival function *S(t)* and probability density function *f(t).* The survival function for the population of patients entering the clinical trial is given by

$S_{pop}(t)$ = P[There is no detectable cancer by time t]

= P[N=0] + $P[t_1>t, t_2>t, \ldots\ldots, t_N>t \mid N\geq 1]$

$$= \exp(-\theta) + \sum_{N=1}^{\infty} (S(t))^N \frac{\exp(-\theta)\theta^N}{N!}$$

$$= \exp(-\theta F(t)) \qquad (2.2.1)$$

The cure rate is given by $S(\infty) = \exp(-\theta)$.

The density function corresponding to (2.2.1) is

$$f_{pop}(t) = \theta f(t)\exp(-\theta F(t)) \qquad (2.2.2)$$

The hazard function for the population is given by

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)}$$

$$= \theta f(t) \qquad (2.2.3)$$

The hazard function $h_{pop}(t)$ is multiplicative in θ and f and thus has a PH structure, a desirable property in survival analysis whereas the hazard function given by (2.1.3) does not possess the PH structure.

The BCH model, also known as the promotion time cure model or the parametric cure rate model, can be written as a mixture model. Equation (2.2.1) can be written as

$S_{pop}(t) = \exp(-\theta) + \exp(-\theta)(\exp(S(t)\theta)-1)$

$$= \exp(-\theta) + (1-\exp(-\theta)) \frac{\exp(-\theta F(y)) - \exp(-\theta)}{(1-\exp(-\theta))} \qquad (2.2.4)$$

The survival function for the non-cured population is given by

$S^*(t) = P[T>t \mid N \geq 1]$

$$= \frac{\exp(-\theta F(y)) - \exp(-\theta)}{(1-\exp(-\theta))} \qquad (2.2.5)$$

Thus every model defined by (2.2.1) can be expressed as a mixture model with cure rate $\phi = \exp(-\theta)$. In the presence of covariates, the canonical link function $\theta = \exp(X^1\beta)$, where X and β are the covariate and the regression coefficient vectors of dimension k×1 respectively, can be used. The literature existing on BCH model is mainly in the Bayesian context since the population survival function is improper and contains the parameter θ. The parametric and semi-parametric methods of estimation are discussed elaborately in chapter 5 of the book "Bayesian Survival Analysis" by Ibrahim, Chen and Sinha (2001).

Yin and Ibrahim (2005) developed a general class of cure models through Box-Cox (1964) transformation on the population survival function. The authors observed that this family contains the earlier two models as special cases. The model is given by

$(S_{pop}(t|X_i,Z_i)^{\boldsymbol{a}}-1)/\boldsymbol{a} = -\theta(\boldsymbol{a},Z_i)F(t|X_i)$, $\boldsymbol{a} \in [0,1]$,

where X_i and Z_i are covariate vectors corresponding to the i^{th} individual and $\boldsymbol{a}$ is the transformation parameter. A discrete uniform prior is taken for $\boldsymbol{a}$**.** The authors have discussed the method of estimation and have observed that when $\boldsymbol{a} = 1$, the model reduces to mixture cure rate model and when $\boldsymbol{a} = 0$, the model becomes BCH model.

The work carried out based on mixture model is using frequentist approach whereas the work based on BCH and the general class of cure models is in the Bayesian context. Techniques for estimation of cure rates when there are partially observed or missing covariates have been discussed by Cho, Schenker, Taylor and Zhuang (2001) and Chen and Ibrahim (2001). There is literature based on comparison of cure rates in various groups. The work by Gray and Tsiatis (1989), Sposto, Sather and Baker (1992), Lee and Sather (1995), Broet, et.al. (2001) are a few to be mentioned.

3. COPULAS:

The interest in modeling multivariate survival data is increasing rapidly. For example, in a study involving diabetic patients, the interest may be in studying the times to primary and secondary complications of diabetes in diabetic patients. The two commonly used approaches in modeling multivariate survival data are the random effects (frailty) approach and the marginal approach. In the frailty approach, independence is assumed conditional on a scalar non-negative random variable known as frailty which multiplies the hazard and when mixed over the distribution produces dependence. Estimation of the dependence structure is of primary concern. The marginal distributions are treated as nuisance functions. The marginal approach considers the marginal distributions to be modeled first and then imposes a dependence structure. The major concern here is consistency of the estimators of the parameters and therefore the association among dependent failure times is treated as nuisance. Modeling correlated multivariate survival data using the marginal approach can be easily done using the copulas.

Copula is a Latin word which means connecting or joining together. It is a function which connects the multivariate probability distribution to the univariate probability distribution. If $T_1,T_2,\ldots\ldots,T_n$ are random variables with a joint distribution $F(t_1,t_2,\ldots\ldots,t_n)$ and marginal distributions $F_1(t_1),F_2(t_2),\ldots\ldots\ldots,F_n(t_n)$, then we have

$F(t_1,t_2,\ldots\ldots,t_n) = C(F_1(t_1),F_2(t_2),\ldots\ldots\ldots,F_n(t_n))$

where C denotes a copula function which generates an n-variate distribution function from an arbitrary set of n univariate distributions. It is a multivariate cumulative distribution function defined on the n-dimensional unit cube $[0,1]^n$ such that every marginal distribution is uniform on [0,1].

For a bivariate set-up, we have

$C(u,0) = C(0,v) = 0.$

$C(u,1) = u$

$C(1,v) = v.$

The Frechet's(1951) bounds for all copulas is given by (M(x,y), W(x,y)) where

M(x,y) = max(0,x+y-1) represents perfect negative correlation between the random variables and W(x,y) = min(x,y) represents perfect positive correlation between the random variables. The extension of these properties to n-dimensional copulas and a brief review of the various properties and results and also the different families of copulas is found in Kolev, Anjos and Mendes (2006).

The family of copulas that has been used in multivariate survival analysis is the Archimedean class. Consider a bivariate distribution F(x,y). It is said to belong to an Archimedean class of copulas if it can be written in the form

$F(x,y) = \phi^{-1}[\phi(F_1(x)) + \phi(F_2(y))]$

where ϕ is the generator function which satisfies the following properties :

$\phi(1) = 0,$
$\lim_{x\to 0} \phi(x) = \infty,$
$\phi^{1}(x) < 0,$
$\phi^{11}(x) > 0$

The following three models belong to the Archimedean family of copulas :

1. Clayton's (1978) model :

$$C_\theta(u,v) = \begin{cases} \left(u^{1-\theta} + v^{1-\theta} - 1\right)^{1/(1-\theta)} & ,\theta > 1 \\ uv, \theta = 1 \end{cases}$$

2. Frank's (1979) model:

$$C_\kappa(u,v) = \begin{cases} \dfrac{\log\left\{1 - \dfrac{(1-\kappa^u)(1-\kappa^v)}{1-\kappa}\right\}}{\log\kappa} & ,0 < \kappa < 1 \\ uv, \kappa = 1 \end{cases}$$

3. Positive stable model : (Hougaard, 1986a)

$$C_\omega(u,v) = \begin{cases} \exp[-\{-\log u)^{1/\omega} + (-\log v)^{1/\omega}\}^{\omega}], 0 < \omega < 1 \\ uv, \omega = 1 \end{cases}$$

θ, κ, ω are the copula parameters which measure the degree of association between u and v. The range for the parameters shown above account for positive association only.

The models developed by Clayton and Frank consider negative correlation when $\theta < 1$ and $\kappa > 1$ respectively, whereas the positive stable model considers positive correlation only. Some important work on copulas can be found in Genest and MacKay (1986), Marshall and Olkin (1988) ,Oakes (1989), Shih and Louis (1995), Oakes and Wang (2003) and Escarela et.al.(2006).

4. MULTIVARIATE CURE RATE MODELS:

When the interest is in jointly modeling several types of failure time random variables such as times to relapse of a disease and death, times to the detectability of cancer at two or more organs, times to primary and secondary complications of a disease, familial association between various genetic diseases such as breast cancer, diabetes and heart diseases, the multivariate models are used. There is not much literature available on multivariate cure rate models. The important papers found in the classical framework are that of Chatterjee and Shih (2001) and Price and Manatunga (2001). Both these papers were based on standard cure rate models. In the Bayesian context, the work by Chen, Ibrahim and Sinha (2002) and Yin (2005) are noteworthy and are based on the BCH model.

The model proposed by Chatterjee and Shih(2001) is an extension of the univariate cure rate mixture model to a bivariate setting. The model was developed to analyze the correlated survival data when there exists a cured proportion in the study population. Correlated survival data are those which consider the familial association for diseases like breast cancer, diabetes and heart diseases. Let two members from each family be involved in the study. If Y_1, Y_2 be the two random variables taking values either 1 or 0 when the j^{th} individual is susceptible or not ; j = 1,2 , T_j be the age at onset of the disease when the j^{th} individual is susceptible, then the marginal distributions of Y_j and the failure time T_j for the susceptible individuals are given by $1-\phi_j = P(Y_j=1)$ and $S_j(t)=P(T_j \geq t \mid Y_j = 1)$ respectively. A common marginal distribution for the two members of the family is assumed, that is $\phi_1 = \phi_2$ and $S_1(t) = S_2(t)$ for all t. $S_1(t)$ and $S_2(t)$ may be parametric or non-parametric. A dependence structure between the members

of the pair is specified. The first type of association is between susceptibility to the disease among the two individuals in the pair, that is, between Y_1 and Y_2. This association is given by the pair-wise odds ratio parameter given by $\gamma = \frac{p_{11}}{p_{10}} * \frac{p_{00}}{p_{01}}$ where p_{ij}=P(Y_1=i,Y_2=j),i=0,1 and j = 0,1. The pair-wise odds ratio is used to characterize the dependence between binary outcomes. The second type of association is between the failure times of the two susceptible members, that is between T_1 and T_2. The dependency structure between the failure times of the two susceptibles is specified using the copulas. Another assumption made in their approach was that the marginal distribution of the failure time of one susceptible is independent of the susceptibility status of another i.e., $P(T_j \geq t_j \mid Y_j = 1, Y_i, i \neq j) = P(T_j \geq t_j \mid Y_j = 1)$, j = 1,2.
The estimation of the parameters has been explained elaborately by the authors.

Wienke et.al. (2003) proposed a cure rate mixture model to analyse bivariate time-to-event data. Correlated gamma frailty model was used to specify the dependency structure between the failure times of two susceptibles. As remarked by Chatterjee and Shih (2003), the model developed by Wienke, et.al.(2003) is a particular form of that proposed by Chatterjee and Shih (2001). The model proposed by Price and Manatunga (2001) used a frailty to account for the correlation between individuals.

Chen, Ibrahim and Sinha (1999) developed a bivariate cure rate model in the Bayesian context based on the BCH model. It was obtained as follows. Suppose Y_1 and Y_2 denote the time to relapse of cancer and time to death respectively, with cumulative distribution functions $F_1(y_1)$ and $F_2(y_2)$ respectively, then let N_1 and N_2 be the number of tumor cells capable of metastasizing corresponding to the variables Y_1 and Y_2 respectively. It is assumed that N_1 and N_2 are independent and follow Poisson distribution with mean $\Theta_k\omega$, k =1,2. The component ω is a frailty component introduced into the model to induce correlation between $N_1$and N_2.The survival function for the population is given by

$S_{pop}(y_1,y_2|\omega) = \exp\{-\omega[\theta_1 F_1(y_1) + \theta_2 F_2(y_2)]\}$.

This model was later extended by Chen et.al.(2002) to a multivariate set-up. A positive stable distribution was considered for the frailty component. Good discussions on positive stable distribution are found in Hougaard (1986a,b, 1995), Manatunga (1989), Oakes (1994), Samorodnitsky and Taqqu (1994), Lam and Kuk (1997), Qiou, Ravishankar and Dey (1999).

Yin (2005) proposed two forms of cure rate frailty models based on the BCH model to analyse correlated or clustered failure time data in a multivariate set up. The first model was called the promotion time cure rate frailty model in which the population hazard was given by

$\lambda_{pop}(t|Z,W) = \lambda(t)W\exp(-\Lambda(t)W)\exp(\beta^1 Z)$.

The second model was called the cure gamma frailty model for which population hazard was $\lambda_{pop}(t|Z,W) = f(t)W\exp(\beta^1 Z)$.

Z denotes the covariate vector, W denotes the frailty, $\lambda(t)$and $f(t)$ are the unknown and unspecified baseline hazard and density functions respectively, and $\Lambda(t)$ is the cumulative hazard function. The methods of estimation are discussed by the author with reference to an example from dental set up.

There is relatively little work on joint modeling of longitudinal and survival data with a cure fraction. A joint longitudinal and survival cure mixture model was proposed by Law, Taylor and Sandler (2002).The maximum likelihood estimators of the parameters was obtained using Monte Carlo Expectation Maximization (EM) algorithm. Brown and Ibrahim (2003) also proposed a model in the Bayesian context using the BCH model for jointly modeling longitudinal and time-to-event data with a cure fraction. The estimation of the parameters was done using the Gibbs sampler (Gelfand and Smith (1990)).

An interesting point to be noted is that it is futile to discuss about cure rate estimation when there is no cured proportion. The cure models will face the problem of estimation of parameters when there is no cured fraction and therefore the non-parametric methods suggested by Maller and Zhou (1992, 1995) should be applied first to test for the

presence of the cured fraction in the univariate set up. This work needs to be extended to handle multivariate data.

5. CURE MODELS FOR RECURRENT EVENT OR COUNT DATA:

In the previous sections, the main focus of the discussion was on the time to occurrence of an event and the associated cure model. In diseases like epilepsy, asthma and urinary tract infection, the variable of interest is the number of times the incident occurs during the follow up period. Thus the response variable or the variable of interest is the number of recurrences of the event of interest. In such cases, the factors associated with the response variable can be modeled using a discrete distribution. The commonly used model is the generalized linear model based on Poisson distribution. Clayton (1994) has shown the connection between the generalized linear models for the recurrent event data and the Proportional Hazards (PH) model of Cox (1972). However Clayton (1994) does not deal with the cure models for the recurrent event data.

In the present paper, a cure model for recurrent event data is proposed. When the recurrent event data is modeled through Poisson distribution, we observed that the associated cure model is based on the inflated Poisson distribution. To simplify the situation, we assumed that the duration of follow up is the same for all the patients and there are no missing data, that is, the data corresponding to the patients who are lost to follow up. Let Y be the random variable denoting the number of times the event of interest occurs. The probability mass function (p.m.f.) of Y is given by

$$P[Y=y] = \begin{cases} p+(1-p)\exp(-\lambda) , & y=0 \\ (1-p)\exp(-\lambda)\dfrac{\lambda^y}{y!} , & y\neq 0 \end{cases}$$

Here p denotes the proportion of patients cured in the population. The covariates can be linked to the parameter λ or p through the link functions

$$logit\ p_i = \log\left(\frac{p_i}{(1-p_i)}\right) = \alpha_0 + \alpha_1 z_{i1} + \ldots\ldots\ldots\ldots + \alpha_k z_{ik}$$

and

$$log\ \lambda_i = \beta_0 + \beta_1 x_{i1} + \ldots\ldots\ldots\ldots + \beta_p x_{ip}$$

where x_i and z_i are the covariate vectors and p_i and λ_i are parameters associated with the i^{th} individual.

In the recent years, a lot of papers have appeared on the zero inflated Poisson and related distributions like inflated negative binomial and inflated generalized Poisson. Some of the papers include that of Lambert (1992) and Bhattacharya et.al.(2008) and the references cited therein. However the inflated distributions have not been used in cure models. We have used this model to estimate the proportion of people who do not have dental caries (i.e., proportion of people who are insusceptible to the adverse event of interest). A survey was conducted by A.B.Shetty Memorial Institute of Dental Sciences, Karnataka, India, at two of their rural health centres, to find the number of people having dental caries and the factors associated with caries at a given point of time. A total of 2000 people who came for health check up were surveyed. The age, gender, occupation, dietary habits, brushing habits viz., the type, frequency and method, the present and past history of medication of the patients, whether self or prescribed by a physician or a medical practitioner and also their DMFT (Decayed, Missing, Filled Teeth) index were recorded. The DMFT index gives the number of Decayed, Missing, Filled Teeth in an individual. The lower the DMFT index, the better the oral hygiene. An individual is said to have hygienic oral health if his DMFT index is zero.

Zero inflated Poisson distribution was used to estimate the proportion of people who did not have dental caries. In the model, log-link function was used to relate the parameter λ and the inflated parameter p was treated as a constant. The analysis was carried out using STATA 7.5 using backward elimination procedure. The results are summarized in the table below.

Variable	Coefficient	Standard	P - value	95% confidence

	β	Error		interval
age	0.1834	0.0209	< 0.001	(0.1423, 0.2244)
Brushing frequency	-0.2244	0.0457	< 0.001	(-0.3139, -0.1347)
Past history of prescribed medication	-0.4353	0.1366	0.001	(-0.7029, -0.1675)
Self medication	0.4002	0.1354	0.003	(0.1349, 0.6656)

A glance at the table indicates that the variables influencing the oral health status are the age, the brushing frequency, self medication and past history of prescribed medication. The estimate of logit p is -0.8734 with a standard error of 0.057011 which is highly significant with a p-value of less than 0.001. Thus the estimate of the proportion p insusceptible to dental caries is 0.2945 (nearly 30%). In this example, p denotes the proportion of people having the DMFT index as zero. The regression coefficients for the variables age and self medication are positive which is an indication that the DMFT index has direct correlation with these variables. Thus the aging factor and self medication will lead to a higher value of DMFT index which is an indication of deteriorating oral health status in old age and also when an individual takes medicines without consulting medical practitioners . There is negative association between the DMFT index and the variables brushing frequency and past history of medication. The increase in the frequency of brushing will lead to a lower value of DMFT index which shows that the more we brush, the better is our oral health status. The same is the oral health condition when we regularly visit medical practitioners for a health check up. The DMFT index is an example for count data.

The model derived by the authors naturally arises from the BCH model of Yakovlev et.al. (1993) and the mixture model. In the derivation of the model,

$$P[Y=0]=\exp(-\theta)+(1-\exp(-\theta))\exp(-\lambda)$$

$$P[Y=k]=\frac{(1-\exp(-\theta))\exp(-\lambda)\lambda^{k}}{k!},\quad k = 1, 2, 3,\ldots..$$

if we use m(t) as the number of times the event occurs in the cured proportion, then the expression for the number of times the event occurs in the population reduces to

$$P[Y=0]=\exp(-\theta)+(1-\exp(-\theta))\exp(-\lambda)$$

the probability that the event does not recur. The cure rate is p = exp(-θ). Since Poisson distribution is used to determine the number of cells capable of metastasizing, the natural choice for the distribution of m(t) is again Poisson with parameter λ and the model reduces to zero inflated Poisson model. This is the biological explanation for the use of zero inflated Poisson distribution as a cure model for recurrent event or count data.

The example considered in the paper relates to the current status data and the use of cure models may not be fully justified. However, as noted by the U.U.D.M. Project Report 2007:4 of Andersson, the use of cure models provides useful information regarding the proportion of people who may not be suffering from the disease at that point of time. In the present context, the cured proportion provides us the information that nearly 30% of the people did not require dental care at the time of the survey.

6. CONCLUSION :

In this paper, an exposition of the cure rate models for the medical practitioners and researchers who may be interested in using the cure rate models but lack the technical knowledge to follow that which have appeared in various statistical journals, is given. Also, cure rate models for recurrent event or count data which is an extension of the work by Clayton (1994), is given. A biological explanation for this model is also given using the BCH model and is a natural extension of the cure rate model developed by Yakovlev et.al. (1993). An application of the model to the data from a dental health set up provides good information regarding the percentage of population who do not suffer from any dental problems at a given point of time. Additional information could have been gathered if the covariates were simultaneously used for modeling the cure fraction p and the regression model for λ but due to the convergence problem in the estimation of the parameters, it could not be done.

7. AREAS OF FUTURE RESEARCH:

Looking at the previous work from our literature survey, we identify the following areas of potential research.

- The development of a formal test for testing for sufficient follow-up of subjects in a heterogeneous population with correlated failure time data.
- The extension of the model proposed by Chatterjee and Shih (2001) to a multivariate set-up.
- The development of regression diagnostics for joint modeling of longitudinal and survival data with a cure fraction.
- The treatment of the missing data for the discrete cure models. The work by Little and Rubin (1987) seems to be promising but needs careful research in the applications relating to censored data involving counts.
- The extension of the cure models for the correlated recurrent event data. An answer to this question is the use of multivariate inflated Poisson distribution (Gan (1999), Li, et.al.(1999)). However the ML estimation is computationally tedious and intractable using the present softwares. Research in this direction is called for.

BIBLIOGRAPHY:

Andersson ,T.(2007).Analyzing time trends in cancer patient survival using cure fraction models. *U.U.D.M. Project Report* 2007: 4, *Department of mathematics, Uppsala University* .

Angelis, R.D., Capocaccia, R., Hakulinen, T., Soderman, B. and Verdecchia, A. (1999). Mixture models for cancer survival analysis : Application to population based data with covariates. *Statistics in medicine* **18**, 441-454.

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

Betensky, R. A. and Schoenfeld, D.A. (2001). Nonparametric Estimation in a cure model with Random cure times. *Biometrics* **57**, 282-286.

Bhattacharya, A., Clarke, B.S., and Datta, G.S. (2008). A Bayesian test for excess zeros in a zero-inflated power series distribution. *IMS collections, Beyond Parametrics in Interdisciplinary Research : Festschrift in Honor of Professor Pranab K Sen*, **1,** 89-104.

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

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26**, 211-252.

Broet,P., Rycke, Y.D., Tubert-Bitter, P., Lellouch, J., Asselain, B ., and Moreau, T. (2001) . A Semiparametric approach for the two-sample comparison of survival times with long-term survivors .*Biometrics* **57**, 844-852.

Brown,E.R. and Ibrahim,J.G. (2003). Bayesian approaches to joint cure rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59,** 686-693.

Chatterjee, N. and Shih ,J.H. (2003). On use of Bivarate Survival models with cure fraction. *Biometrics* **59**, 1184-1185

Chatterjee,N. and Shih,J. (2001). A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics* **57**, 779-786.

Chen,M.H., Ibrahim,J.G., and Sinha,D. (2002). Bayesian inference for multivariate survival data with a cure fraction. *Journal of Multivariate Analysis* **80**, 101-126

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

Chen, M.H and Ibrahim ,J.G.(2001) . Maximum likelihood methods for cure rate models with missing covariates .*Biometrics* **57,** 43-52

Cho,M., Schenker,N., Taylor, J.M.G., and Zhuang, D (2001). Survival analysis with long-term survivors and partially observed covariates .*The Canadian Journal of Statistics* **29,** 421-436

Clayton,D.G.(1994). Some Approaches to the analysis of recurrent event data. *Statistical methods in medical research* **3**,244-262

Clayton,D.G. (1978). A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141-151.

Cox,D.R (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Escarela , G., Mena, R.H., and Castillo-Morales ,A .(2006). A Flexible class of parametric transition regression models based on copulas : application to poliomyelitis incidence. *Statistical methods in medical research* **15**,593-609.

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

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

Frank,M.J. (1979). On the simultaneous associativity of F(x,y) and x+y-F(x,y). *Aequationes Mathematicae* **19**, 194-226.

Fréchet, M. (1951). Sur les Tableaux de Corrēlation Don't les Marges Sont Donnēes. *Annales de l'Universite' de Lyon Série* 3, **14**, 53-77

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

Gan, N. (1999). General zero inflated models and their applications. *Unpublished Ph.D. thesis submitted to North Carolina State University.*

Gelfand,A.E. and Smith,A.F.M.(1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Genest,C., and Mackay,R.J. (1986). The joy of copulas : bivariate distributions with uniform marginal. *The American Statistician* **40**, 280-283.

Ghitany,M.E.,Maller,R.A., and Zhou,S. (1994) - Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis* **49**, 218-241.

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

Gordon,N.H. (1990). Application of the theory of finite mixtures for the estimation of cure rates of treated cancer patients. *Statistics in Medicine* **9**, 397-407.

Gray, R.J. and Tsiatis, A. A. (1989). A linear rank test for use when the main interest is in differences in cure rate. *Biometrics* **45**, 899-904.

Greenhouse,J.B. and Wolfe,R.A.(1984). A competing risks derivation of a mixture model for the analysis of survival. *Communications in Statistics – Theory and methods* **13**, 3133-3154.

Haybittle, J.L.(1965). A two parameter model for the survival curve of treated cancer patients. *Journal of the American Statistical Association* **53**, 1, 16-26.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime data Analysis* **1**, 255-273.

Hougaard P (1986 a). A Class of Multivariate Failure Time Distributions. *Biometrika* **82,** 543 - 552

Hougaard, P. (1986 b). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73**, 387-396, (Correction, **75** 395).

Ibrahim, J.G., Chen, M.H. and Sinha, D. (2001). Bayesian Survival Analysis. *Springer-Verlag New York, Inc.*

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481.

Kolev, N., Anjos , U., and Mendes ,B.V.M.(2006). Copulas : A Review and recent developments. *Stochastic models* **22**,617-660.

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

Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics* **34**, 1-13.

Lam, K.F. and Kuk, A.Y.C.(1997). A marginal likelihood approach to estimation in frailty models. *Journal of the American Statistical Association* **92**, 985-990.

Laska,E.M., and Meisner,M.J. (1992). Nonparametric estimation and testing in a cure rate model. *Biometrics* **48**, 1223-1234.

Law,N.J., Taylor,J.M.G., and Sandler,H. (2002). The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics* **3,** 547-563.

Lee,J.W. and Sather ,H.N. (1995) .Group Sequential methods for comparison of cure rates in Clinical trials. *Biometrics* **51**,756-763.

Li, C.S., Lu, J.C., Park, J., Kim, K., Brinkley, P.A. and Peterson, J.P. (1999). Multivariate zero inflated Poisson models and their applications. *Technometrics* **41**, 29-38.

Little, R.J.A. and Rubin, D.B (1987) . *Statistical Analysis with Missing Data* .New York: Wiley.

Maller, R. A. and Zhou, S. (1996) . *Survival Analysis with Long-Term Survivors.* New York: Wiley.

Maller, R. A. and Zhou, S. (1995). Testing for the presence of immune or cured individuals in censored survival data .*Biometrics* **51,** 1197-1205.

Maller ,R .A and Zhou,S. (1994). Testing for sufficient follow-up and outliers in survival data . *Journal of the American Statistical Association* **89,** 1499-1506.

Maller ,R .A and Zhou,S. (1992). Estimating the proportion of Immunes in a censored sample. *Biometrika* **79,** 731-739.

Manatunga, A.K. (1989) .Inference for multivariate survival distributions generated by stable frailties. *Unpublished Ph.D. Dissertation*.Department of Biostatistics, University of Rochester, Rochester, New York .

Marshall, .A.W and Olkin, .I (1988). Families of Multivariate Distributions. *Journal of the American Statistical Association,* **83**, 834 – 841

Mendenhall , W. and Hader, R.J. (1958). Estimation of Parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika* **45** ,504-520.

Miller, R.G.(1981). *Survival Analysis* ,Chichester, U.K : Wiley .

Oakes, D. and Wang,A (2003). Copula model generated by Dabrowska's association measure. *Biometrika* **90**,478-481

Oakes, D. (1994). Use of frailty models for multivariate survival data. *Proceedings of the XVII th International Biometrics conference,* Hamilton, Ontario, Canada, 275-286.

Oakes,D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**, 487-493.

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

Peng,Y., Dear,K.B.G., and Denham,J.W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine* **17**, 813-830.

Price,D.L., and Manatunga,A.K. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine* **20**, 1515-1527.

Qiou, Z., Ravishanker, N., and Dey ,D.K. (1999) .Multivariate survival analysis with positive frailties . *Biometrics* **55**, 637-644.

Samorodnitsky, G. and Taqqu, M.S (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, London : Chapman and Hall.

Shih ,J.A . and Loius,T.A.(1995). Inferences on the Association Parameter in Copula Models for Bivariate Survival Data. *Biometrics* **51**, 1384-1399.

Sposto.R., Sather, H.N. and Baker,.S.A (1992). A comparison of tests of the difference in the proportion of patients who are cured. *Biometrics***, 48**, 87 – 99.

Sy, J.P. and Taylor, J.M.G. (2000). Estimation in a proportional hazards cure model. *Biometrics* **56,** 227-336.

Taylor J.M.G. & Kim D.K. (1993). Statistical Models for Analysing Time to Occurrence Data in Radio Biology and Radio Oncology. *International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine***. 64**, 627 – 640.

Taylor ,J.M.G (1995). Semi-parametric estimation in failure time mixture models. *Biometrics* **51,** 899-907.

Wienke,A., Lichtenstein , P., and Yashin ,A.I (2003). A Bivariate frailty model with a cure fraction for modeling familial correlations in diseases .*Biometrics* **59**, 1178-1183.

Yamaguchi,K.(1992). Accelerated failure-time regression models with a regression model of surviving fraction : An Application to the analysis of " permanent employment" in Japan . *Journal of the American Statistical Association* **87** ,284 – 292.

Yakovlev A.Y. and Tsodikov .A.D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. New Jersey: World Scientific.

Yakovlev A.Y. (1994). Letter to the Editor. *Statistics in Medicine* **13**, 983 – 986

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

Yin, G. (2005). Bayesian Cure Rate Frailty Models with Applications to a Root Canal Therapy Study. *Biometrics* **61,** 552 – 558.

Yin, G and Ibrahim , J.G. (2005). Cure rate models : a unified approach .*The Canadian Journal of statistics* **33**, 559-570