Application of Modified Generalized–Gamma Mixture Cure Model in the Analysis of Ovarian Cancer Data

Serifat A. FOLORUNSO1*, Timothy A.O. OLUWASOLA2, Angela U. CHUKWU1, and Akintunde A. ODUKOGBE2

1Department of Statistics, University of Ibadan, Nigeria
2Department of Obstetrics and Gynecology, College of Medicine, University of Ibadan, Ibadan, Nigeria

*Email: serifatf005@gmail.com

Abstract. The modeling and analysis of lifetime for terminal diseases such as cancer is a significant aspect of statistical work. This study considered data from thirty-seven women diagnosed with Ovarian Cancer and hospitalized for care at the Department of Obstetrics and Gynecology, University of Ibadan, Nigeria. Focus was on the application of a parametric mixture cure model that can handle skewness associated with survival data – a modified generalized-gamma mixture cure model (MGGMCM). The effectiveness of MGGMCM was compared with existing parametric mixture cure models using Akaike Information Criterion, median time-to-cure and variance of the cure rate. It was observed that the MGGMCM is an improved parametric model for the mixture cure model.

Keywords: cure fraction, mixture cure model, modified generalized-gamma, ovarian cancer, parametric, skewness, survival analysis.

1. Introduction

The evaluation of cure fractions in oncology research under the well-known cure rate models had attracted considerable attention while the benefits of these cure rate models over the traditional methods of survival analysis, including the well-known Cox regression model, have been widely investigated [1,2,3]. However, in certain types of cancers such as breast cancer and leukemia, a significant fraction of patients, called the cured proportion or immune or long-term survivors, may now be ‘cured’ through therapy.

The population of interest can therefore be divided into two: cured and non-cured with cure rate models providing satisfactory models in such cases [4]. The survival figures obtained are utilized in choosing treatment types, regimens, doses as well as in discriminating between side effects profiles and cost effectiveness. Thus, cure models are survival models which were developed to estimate the proportion of patients cured in clinical trials as they provide more information about patient survival, predict the proportion of those cured from cancer, predict time until cure and also estimate survival time for those not cured.
The Mixture cure models are useful in evaluating the percentage of patients cured and the survival function of the uncured patients [5,6]. Some authors such as [7,8] developed an alternative model to the mixture cure model. This model is known as the Bounded Cumulative Hazard (BCH) model which is also denoted as the non-mixture cure model. It is therefore appropriate to consider the fact that the proportion cured and the mean survival time for the non-cured patients can be useful summary parameters for detailed assessment of the differences in survival. Exploring some existing cure models like Lognormal Mixture Cure Model, Loglogistic Mixture Cure Model, Weibull Mixture Cure Model, and Generalised-Gamma Mixture Cure Model (GGMCM). In a study by [9], it was established that GGMCM out-performed other parametric cure models in terms of Akaike Information Criterion (AIC) but could not handle acute asymmetry in survival data. The main aim of this study was to apply a mixture cure model that is capable of handling and accommodating non-normality in survival data in order to obtain better information of the proportion of ovarian cancer patients that have benefited from medical intervention in comparison with existing models.

2. Methodology
This study adopted a retrospective, descriptive, non-intervention approach in evaluating a data set retrieved from the records of patients managed for ovarian cancer at the Gynecology Oncology Unit of the University College Hospital (UCH) Ibadan, Nigeria. UCH is the foremost tertiary centre for cancer care in Nigeria. The hospital receives referrals from all states in Nigeria and beyond thus making it the hospital of choice for a large proportion of cancer patients. All cases of ovarian cancer diagnosed between July 2004 and April 2015 were followed up till February 2017 in order to determine the individual’s survival duration. The survival time was calculated from the date of diagnosis to the date of death or termination of follow up. The follow-up of patients was by personal contacts in the oncology unit and phone calls as necessary. In addition, relations were also contacted and interviewed as appropriate so as to determine the vital status of patients in cases of missing or unclear details from available records.

The ethical considerations of beneficence, non-maleficence, equity and confidentiality were strictly applied and adhered to in handling the patient’s data. Serial numbers were used during data entry, analysis and reporting so as to ensure anonymity. Ethical approval was obtained from the joint University of Ibadan and University College Hospital, Ibadan ethics committee.

The study explored the Mixture Cure Model (MCM), Boag (1949) was the first author that developed this model and it was revised by [6] as:

$$S(t) = c + (1 - c)S_u(t)$$

(1)

Where
- $S(t)$ = Population Survival function
- $S_u(t)$ = Uncured patients Survival function
- $c$ = Proportion cured

In a mixture cure model, the population involves two components where those with probability $P(1 - c)$ are the uncured component and $P(c)$ the cure component. A technique used for parameter estimation in this study is Maximum Likelihood Estimation (MLE) method has previously by some authors [10,11,12]. Conventionally, estimations of mixture cure model parameters can either follow parametric or non-parametric techniques. The present study goes through a parametric system of estimation using the link function from gamma family [13]. Considering the given equation (1) above, parameter $c$ of the mixture cure model can be acquired as follow: the cure rate was incorporated into the likelihood function of MCM, using the likelihood function given below in Equation 3:
\[ L = \prod_{i=1}^{n} [f(t_i)]^{d_i} [S(t_i)]^{1-d_i} \]  

(2)

Where \( d_i = \begin{cases} 
0, & \text{if patient is cured} \\
1, & \text{otherwise} 
\end{cases} \)

\( d_i \) is 0 if the patient is cured and 1 if the patients is not cured.

\( f(t) = \text{pdf of parametric cure model} \)

\( S(t_i) = \text{survival function} \)

Recall that \( S(t) = 1 - F(t) \), hence equation (1) becomes;

\[ 1 - F(t) = c + (1 - c)(1 - F_u(t)) \]  

(3)

Differentiating (3); its becomes

\[ f(t) = \left( (1 - c)f_u(t) \right) \]  

(4)

Put (4) in equation (2) to obtain

\[ L = \prod_{i=1}^{n} \left[ (1-c)f_u(t_i) \right]^{d_i} \left[ c + (1-c)S_u(t_i) \right]^{1-d_i} \]  

(5)

The log-likelihood function of mixture cure model can be gotten with the aid of taking logarithms of equation (5) as follow:

\[ \log L = \sum_{i=1}^{n} d_i (1 - c) + d_i \log f_u(t_i) + \sum_{i=1}^{n} (1 - d_i) \log \left[ c + (1-c)S_u(t_i) \right] \]  

(6)

The study further developed an innovative cure model termed modified GGMCM. The following pdf and cdf were used jointly to achieve the proposed model. Following the assumption that random variable \( t \) follows a generalised-gamma distribution, and then its density function is defined follows as:

\[ f(t) = \frac{\beta}{\theta \Gamma\alpha} \left( \frac{t}{\theta} \right)^{\beta-1} e^{-\left( \frac{t}{\theta} \right)^{\beta}} \]  

(7)

and its cdf as;

\[ F(t) = \frac{\gamma \left[ \alpha \left( \frac{t}{\theta} \right)^{\beta} \right]}{\gamma(\alpha)} \]  

(8)

But in scale location form, equations (7 & 8) becomes

\[ f(t) = \frac{1}{\sigma \Gamma\alpha} e^{-\left( \frac{\log t - \mu}{\sigma} \right)} e^{-\left( \frac{t - \mu}{\sigma} \right)} \]  

(9)

and

\[ F(t) = \frac{\gamma \left[ \alpha, e^{-\left( \frac{\log t - \mu}{\sigma} \right)} \right]}{\Gamma(\alpha)} \]  

(10)

If we put the pdf and cdf in equation (9 & 10) into (7), we have
\[ g(t) = \frac{1}{\Gamma \beta} \left[ -\log \left( 1 - \frac{\gamma \left[ \alpha, \frac{\log t - \mu}{\sigma} \right]}{\Gamma \alpha} \right) \right]^{\omega-1} \frac{1}{\sigma \Gamma \alpha} e^{-\frac{(\log t - \mu)}{\sigma}} - e^{-\frac{\log t - \mu}{\sigma}} \]  

(11)

Where in the above equation (8), the parameter \( \beta \) is equal to the parameter \( \alpha \) in the same equation, thus, cdf of gamma generalised link function is

\[ G(t) = \frac{\gamma(-\log(F(t)), \beta)}{\Gamma \beta} \]  

(12)

Putting equation (10) in scale location form through equation (12), we obtain the cdf of modified generalised Gamma as

\[ G(t) = \frac{\gamma \left[ \alpha, \frac{\log t - \mu}{\sigma} \right]}{\Gamma \beta} \]  

(13)

Equations (11) and (13) above provide a new model of parametric mixture cure fraction. This is good for the management of acute asymmetry in survival data. They are the density function (pdf) and cumulative distribution function (cdf) of the modified GGMCM. Using the Known function of density and survival from Modified-Generalised Gamma Mixture Cure Models and plugging them into log likelihood functions from cure models in equation (6), it gives equation (14) below:

\[ \log L = \sum d_i \log(1 - C) + \Sigma d_i \log g(t_i) + \Sigma(1 - d_i) \log[C + (1 - C)S(t_i)] + \Sigma(1 - d_i) \log[C + (1 - C)S(t_i)] \]  

The differentiation of equation (7) for \( c \) and solve the equation to zero provides the likelihood equation as
\[
\frac{\partial \log L}{\partial c} = - \frac{\Sigma d_i}{1-c} + \frac{\Sigma (1-d_i)}{1-c} \left[ 1 - \gamma \left( \frac{\alpha \left( \log \frac{t-u}{\sigma} \right)}{\Gamma(\alpha)} \right), \beta \right] \]
\[
\frac{\partial^2 \log L}{\partial c^2} = - \frac{\Sigma d_i}{(1-c)^2} + \frac{\partial}{\partial c} \left[ \frac{\Sigma (1-d_i)}{1-c} \left[ 1 - \gamma \left( \frac{\alpha \left( \log \frac{t-u}{\sigma} \right)}{\Gamma(\alpha)} \right), \beta \right] \right] \]

Furthermore, the second derivatives equation (15) above for c is

\[
\frac{\partial^2 \log L}{\partial c^2} = - \frac{\Sigma d_i}{(1-c)^2} + \frac{\partial}{\partial c} \left[ \frac{\Sigma (1-d_i)}{1-c} \left[ 1 - \gamma \left( \frac{\alpha \left( \log \frac{t-u}{\sigma} \right)}{\Gamma(\alpha)} \right), \beta \right] \right] \]

It is not possible to achieve equation (11) in a close form after having equated it to zero when resolving for c, it is therefore resolved through numerical repetitive process.

3. Analysis and results

The summary statistics depicting the descriptive information of the 37 life ovarian cancer data used in this study and their survival time is presented in table 1, table 2 showed the model selection criterion used in this study for evaluating and validating the modified generalized-gamma mixture cure model and table 3 shows the most important benefit of using confidence interval in Statistical analysis. The scatter plot for the survival period in months using the observations of ovarian cancer patients was shown in Figure 1a; this plot describes the ovarian cancer data. The minimum and maximum numbers of month were 1 and 159 respectively. The plot in Figure 1b is survival period histogram plot of the ovarian cancer data for all the observed patients. The plot describes and gives pictorial representation of the distribution of ovarian cancer data. The plot is skewed to the right which also supports the literature about survival data being rightly skewed. Figure 1c showed the survival period normal QQ Plot of the ovarian cancer data for all the observed patients. This plot describes the quantiles of two distributions against each other and those patterns can be equally used for comparison of two distributions such that data with outliers are detectable in the upper right corner of the plot. The Survival Period Density Plot of the ovarian cancer data for all the observed patients is presented in figure 1d. The plot in Figure 1e contains Survival Period Box-Plot of the ovarian cancer data for all the observed patients and provides a visual representation of a distribution; the whiskers are the two lines outside the box, which extends to the smallest and largest observation of the ovarian cancer data except outliers which are shown as dots outside the whiskers. Finally, Figure 1f contains survival period of cumulative distribution function (cdf) Plot of the ovarian cancer data as it converges to 1. Figure 2(a-f) shows the Density function plots of some of the existing parametric distributions as well as the modified GGCM. It comprises Density function plots of Weibull mixture cure model, log-logistics mixture cure model, lognormal mixture cure model, Generalised-Gamma Mixture cure model and modified Generalised-Gamma Mixture cure. Figure 3(a-e) comprises their respective Survival and Hazard function plots while Figure 4 shows the combined cumulative density plots of all the models with the modified GGCM converging to 1 faster than the existing models. The Exploratory Data Analysis showed the minimum and
maximum observed data, measures of locations like average and median as well as measures of spread such as quartiles. The summarization included other statistics that can show description of the ovarian cancer data such as skewness and kurtosis which are measures of symmetry. There is an outlier event at observation 37 (159 months) which supports the fact that survival data are positively skewed in nature. The smoothening of the distribution of points alongside the axis can be showcased using Survival Period Density Plot.

Table 1. Descriptive statistics of the survival time of ovarian cancer

| Min | Q1 | Mean | Median | Q3 | Max | Skewness | Kurtosis |
|-----|----|------|--------|----|-----|----------|----------|
| 1month | 24months | 48months | 45.65months | 57months | 159months | 1.395 | 7.339 |

Table 2. Model evaluation of the ovarian cancer data

| Model | AIC | -2loglike | µ̅ | σ̅ | eµ̅ | c | Var(c) |
|-------|-----|-----------|----|----|-----|---|-------|
| Weibull | 216.88450 | 210.88460 | 2.44000 | 68.2000 | 60.07060 | 0.28900 | 0.0968100 |
| Lognor | 205.97820 | 199.97820 | 4.04000 | 0.32800 | 57.97480 | 0.37120 | 0.0042610 |
| Llog | 203.27440 | 197.27440 | 5.97010 | 56.2070 | 56.86490 | 0.18100 | 0.2878700 |
| GG | 206.20140 | 198.20140 | 3.96350 | 0.30170 | 20.90111 | 0.10840 | 0.0052300 |
| MGG | 199.23530 | 194.47120 | 7.52400 | 15.2460 | 11.82010 | 0.82070 | 0.0001253 |

Table 3. Confidence interval of cure fraction model (c)

| Model | C | Var(c) | SE(c) | SE(c) × Zα/2 | c ± Zα/2SE(c) | CI |
|-------|---|--------|-------|--------------|----------------|----|
| Weibull | 0.28900 | 0.0986100 | 0.3111431 | 0.098405 | 0.28900 ± 0.6098405 | [-0.3208405, 0.8988405] |
| Lognormal | 0.37120 | 0.0042610 | 0.0652763 | 0.1279415 | 0.3712 ± 0.1279415 | [0.2432585, 0.4991415] |
| Llog | 0.18100 | 0.2878700 | 0.5365352 | 1.05160900 | 0.18100 ± 1.05160900 | [-0.870609, 1.232609] |
| GG | 0.10840 | 0.0052300 | 0.0723187 | 0.1417447 | 0.10840 ± 0.1417447 | [-0.0333447, 0.2501447] |
| MGG | 0.82070 | 0.0001253 | 0.0111938 | 0.0219398 | 0.82070 ± 0.0219398 | [0.7987602, 0.8426398] |
Figure 1. Exploratory data analysis (Ovarian Cancer)
4. Discussion
The study employed the following criteria to determine the model efficiency: Akaike Information Criterion, loglikelihood, variances, cure rate (c), median time-to-cure, variance of cure rate and median survival time. The highest concentration points can be found in the peak of the density. In the real sense, density plots tend to be quite reliable and informative for large data sets but can be misleading for data sets of only a few points. The survival period of the cdf indicated that the modified GGMCM was able to estimate the ovarian cancer data better as it converges to 1 faster than the existing models. This is similar to the report of [14] when a mixture cure fraction model with random effects was applied to cause-specific survival data of female breast cancer patients obtained from the population-based Finnish Cancer Registry. The authors used two sets of random effects to capture the regional variation in the cure fraction and in the survival of the non-cured patients. Therefore, the confidence interval above gives the 95% confidence interval of the point estimate ‘c’ for all the considered models. It is clearly observed that there is close margin of c in the proposed model. The margin of error associated with the proposed model MGGMCM is very small compared to the conventional models. The confidence interval of cure fraction with a very low margin of error gives a better one.

From the analysis of the data, it was revealed that all the model criteria used for this study agreed with the underlying assumptions that existing models were significantly less effective than the modified Gamma-Generalized Mixture Cure Model (MGGMCM). This showed that the MGGMCM enhanced an improvement and gives a better fit for the data used in comparison with the previous models thus implying that MGGMCM provided lowest criteria. It is therefore established that the MGGMCM is effective in accomplishing more reliable results that can adequately handle the problem of acute-asymmetry associated with survival data.

5. Conclusion
This study has been able to confirm that MGGMCM is the flexible best model that explained the ovarian cancer data used for the study in terms of AIC, value of c and Median time to cure. The MGGMCM can be used effectively to model a sizable number of data sets as the results have been able to show that it is an improved model for statistical modeling and inference for survival data that exhibits high degree of asymmetry.
References
[1] Hsu, Wei-Wen Todem, David and Kim KyungMann. (2016). A Sup-Score Test for the Cure Fraction in Mixture Models for Long-Term Survivors. *Biometrics* DOI: 10.1111/biom.12514
[2] Sposto R. (2002) Cure model analysis in cancer: an application to data from the Children? Cancer Group. *Stat Med.*, 21: 293- 312.
[3] Goetghebeur, E. and L. Ryan, (2000). Semi-parametric regression analysis of interval-censored data. *Biometrics*, 56: 1139-1144.
[4] Othu, M., Y. Li and R. Tiwari, (2012). Change-point cure models with application to estimating the change-point effect of age of diagnosis among prostate cancer patients. *J. Applied Stat.*, 39: 901-911
[5] 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
[6] Berkson J. and Gage R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501-515.
[7] Yakovlev, A.Y., Tsodikov, A.D. & Bass, L., (1993). A stochastic-model of hormesis. *Mathematical Biosciences* 116 (2), 197-219.
[8] Tsodikov, A. D., Ibrahim, J. G., and Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *J. Amer. Statist. Assoc.*, 98(464), 1063 - 1078.
[9] Chukwu A.U and Folorunso S.A (2015). Determinant of flexible Parametric Estimation of Mixture Cure Fraction Model: An Application of Gastric cancer Data. West African Journal of Industrial & academic research Vol. 15 No.1 : 139- 156.
[10] De Angelis R, Capocaccia R, Hakulin T, Soderman B, Verdecchia A. (1999), Mixture models for cancer survival analysis: application to population-based data with covariates. Statistics in Medicine; 18:441
[11] Lambert PC, Thompson JR, Weston CL, Dickman PW. (2007) Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics* 2007; 8:576- 594. DOI: 10.1093/biostatistics/kxl030.
[12] Verdecchia A, De Angelis R, Capocaccia R, Sant M, Micheli A, Gatta G, Berrino F (1998). The cure for colon cancer: results from the EUROCASE study. *International Journal of Cancer*; 7:322-329.
[13] Zografos, K. and Balakrishnan, N. (2009). On families of beta- and generalized gamma generated distributions and associated inference. *Stat. Method.*, 6, 344-362.
[14] Seppa, K., Hakulinen, H.J. Kim and E. Laara, (2010). Cure fraction model with random effects for regional variation in cancer survival. *Stat. Med.*, 29: 2781-2793.