NONPARAMETRIC ESTIMATE BASED ON IMPUTATIONS TECHNIQUES FOR INTERVAL AND PARTLY INTERVAL CENSORED DATA

Abdallah Zyoud[1,*], F. A. M. Elfaki, and Meftah Hrairi

[1]Department of Mechanical Engineering, Faculty of Engineering, IIUM, P.O.Box 10, 50728 Kuala Lumpur, Malaysia

*E-mail:Abdallahzyoud@yahoo.com

ABSTRACT: This paper discusses the nonparametric analysis of interval and partly interval censored data, which occurs in many fields including demographical, epidemiological, financial, medical and sociological studies. For Partly Interval Censored (PIC) we mean that for some subjects the exact failure time is observed while the rest are only known to fall within an interval [17]. In medical and reliability studies the most important function is the survival function. However, we used a nonparametric model to estimate the survival function based on different imputation techniques in the present of interval censored and PIC data. Our proposal is easily implemented using R software.

Keywords: Nonparametric Estimator; PIC Data; Imputation; Turnbull Estimator; Survival Analysis.

1. INTRODUCTION

Comparison of survival functions is one of the main objectives in survival studies. In this paper, we will discuss the comparison problem for several imputation techniques in the existence of an interval and part interval censored (PIC) failure time data. PIC data often occurs in medical and health studies that are followed by periodic follow-up. By PIC data we mean, for some subjects, the exact failure times are observed, but for the remaining subjects, the survival time of interest is observed only to belong to an interval instead of being exactly [14,17,37]. An example of this kind of data is provided by the Framingham Heart Disease Study [23]. Another example reported elsewhere [18] about Fatigue Failure (Crack size data) which may be considered as PIC data. In this article several imputation techniques used to estimate the survival function and compared with the one that obtained by Turnbull based on interval censored and PIC failure time data. In the next two sections some researchers that have addressed the PIC and imputation techniques.

2. PARTLY INTERVAL CENSORED DATA

Compared to the huge amount of statistical methods developed to tackle right censored, left censored and interval censored data, very little work has been done to deal with PIC data. Here are some of the researches that studied the subject: Among the first to approach interval censored data are [25] who tackled PIC data by treating an exact observation as a very small interval and others [32], who derived self-consistency equations and used the Expectation-Maximization (EM algorithm) iterative procedure to estimate the Nonparametric Maximum Likelihood Estimator (NPMLE), and considered a closed interval $[L, R]$, so that exact observations are taken into account. [11] developed an Iterative Convex Minorant algorithm (ICM) which performed faster than Turnbull's based on EM algorithm especially for large sample sizes; see also [12]. In [3,15,16] some others modified the ICM algorithm, where as in [15,16] authors showed that his modified algorithm always converges. Elsewhere [26], some extended case 2 interval censored data to case *k* where the number of monitoring times is random, the intervals are half open and non-censored observations are not considered. [33] generalize the case 2 model so that exact observations are allowed. [34] proved the consistency of the NPMLE of the mixed case interval censoring in Hellinger distance, and also recovered the results of [26] by using preservation theorems for Glivenko-Cantelli classes. [29] investigated the NPMLE of univariate mixed case interval censored data and provided a characterization of the NPMLE, then used the ICM algorithm to compute the NPMLE. In [27] authors used a method based on a pseudo likelihood ratio for estimating the distribution function of the survival time in a mixed-case interval censoring model and showed that it converges under the null hypothesis to a known limit distribution. [17] was probably the first to use the Cox's Proportional Hazard Model to analyze PIC data. [30] used the Turnbull's self consistency algorithm to determine a nonparametric maximum likelihood estimate of the survival function. Based on which some presented a class of generalized log-rank test for PIC failure time data [37]. Some proposed a Proportional Hazards Weibull Model (PHWM) for PIC data which is parametric Cox model with Weibull distribution and applied it to AIDS studies [5]. Guure *et. al.*[13] tried to determine the best estimate for the Weibull scale parameter using interval-censored survival data. Others compared the performance of five different methods for interval censored data [5]. Finally, others modified the estimating functions for PIC data using the semi-parametric Cox's proportional hazards regression models of the sub-distribution of a competing risks models [6]. In this research, we will tackle PIC data based on several imputation techniques to transform our data into right censored data. The motivation behind that is the imputation process is very simple and there are numerous methods to deal with right censored data.

3. IMPUTATION

Liu, *et. al.* [21] used midpoint imputation to propose a model-based estimate of mean incubation period of AIDS. Mariotto *et. a.l* [22] used midpoint imputation to estimate the acquired immune deficiency syndrome incubation period in intravenous drug users. However, [19] noted that Kaplan-Meier estimates of survival based on the midpoint imputation method may be considerably biased when censoring intervals are wide (longer than two years) and varied.

In [35], authors investigated the effect of infection with GB virus C on survival of patients with HIV infection using right-point imputation. Others [31] also used the right-point imputation method to investigate the effect of infection with GB virus C on the mortality of HIV-infected patients. Some workers [8] used the self-consistency algorithm developed by [32] to estimate G, and then use Ĝ to impute the expected infection time based on conditional mean of the subject's interval. [10] used the multiple imputation method based on Monte Carlo Expectation Maximization algorithm to estimate G, and then repeatedly impute infection times based on random draws from Ĝ conditional on subjects' intervals. [24] proposed a general semi-parametric method based on multiple imputations for Cox regression with interval censored data. [9] compared the conditional mean imputation, multiple imputation, and the midpoint imputation methods for the bias and mean squared error (MSE). [36] compared right-point imputation, midpoint imputation, conditional mean imputation, conditional median imputation, conditional mode imputation, multiple imputation and random imputation methods for doubly censored HIV data. In all previous studies doubly censored data was considered which is most suitable for HIV data where the infection time is interval censored and the time of death is right censored. [4] used multiple imputation method to analyze interval censored data with the additive hazard model. [1] and [2] used multiple imputations for parametric and nonparametric based on partly interval censored data. Finally [20] compared the performance of Cox model with Weibull distribution, right point imputation and mid-point imputation to the illness-death model for interval censored data.

4. METHODS

4.1 SIMPLE IMPUTATION METHODS

There are three main types of simple imputation methods:

1. Right-point imputation where the event time is imputed by the right limit of the interval.
2. Left-point imputation where the event time is imputed by the left limit of the interval.
3. Mid-point imputation which refers to imputing the event time by the midpoint of the interval.

4.2 PROBABILITY-BASED IMPUTATION

4.3 METHODS

Probability-based imputation requires estimating the distribution of the partly interval censored data based on the observed intervals and using our knowledge of the distribution to impute the missing data. Let X be a discrete random variable describing the event time with a asset of values $x = \{x_1, x_2, ..., x_m\}$ associated with a set of probabilities $g = \{g_1, g_2, ..., g_m\}$, respectively, where $x_1 < x_2 < ... < x_m$. Suppose for subject i, there are q_i possible values of failure time $y_i = \{y_{i1}, y_{i2}, ..., y_{iq_i}\} \in [L_i, R_i]$, associated with probabilities $p_i = \{p_{i1}, p_{i2}, ..., p_{iq_i}\}$, $i = 1, 2, ..., n$. Both y_i and p_i are subsets of x and g, respectively. Let $h_i = \{h_{i1}, h_{i2}, ..., h_{iq_i}\}$ be the conditional probability for subject I taking the value y_{ik} is $h_{ik} = \dfrac{p_{ik}}{\sum_{r=1}^{q_i} p_{ir}}$, $k = 1, ..., q_i$ conditioning on the interval $[L_i, R_i]$ and g.

Here are some of the most common probability-based imputation methods:

1. Conditional Mean Imputation where the event time is imputed by the expected value $E(X_i / X_i \in [L_i, R_i], g)$.
2. Conditional Median Imputation where the event time is imputed by the median of y_i weighted by the probability vector h_i. In case the median is not unique $\hat{X}_i$ is taken to be the mean of the two medians.
3. Conditional Mode Imputation where the event time is imputed by the mode of y_i which is the value corresponding to the highest probability. In case the mode is not unique $\hat{X}_i$ is taken to be the average of the modes.
4. Multiple Imputation (MI): MI is one of the most common methods of imputation. For $m = 1, ..., M$, a sample y_{ik}^m is chosen randomly from y_i with replacement using the probability vector h_i as weight. The result of the multiple imputation method will be replacing the original interval censored dataset D with $\hat{D}^m$ which can be easily analyzed with the regular right censored data method, Cox model in our case. Let $\hat{\theta}_m$ be the estimate of the parameter of interest obtained from $\hat{D}^m$, then the MI estimate of θ is $\hat{\theta}_m = \dfrac{1}{M}\sum_{m=1}^{M}\theta_m$.
5. Random Imputation: is a special case of multiple imputation where $M = 1$. Randomly choose one value y_{ik} from the vector y_i using the conditional probability vector h_i as weight.

5. AN EXAMPLE

We applied the proposed method to the modified breast cancer data that was presented by [7]. The data consist of 46 patient treated by Radiation (R) only and 48 patients treated by Radiation plus adjuvant Chemotherapy (R+C). This study was implemented to compare the cosmetic effects of Radiation alone against R+C on women with early breast cancer and the event of interest was the time to first occurrence of breast retraction and the patients were observed at clinic visits every 4 to 6 months, where the actual dates of the event were recorded exactly if available. If not the interval of events were noted. The modified data set is shown in Table 4.1 in order to set up the data as the partly interval censored data, for instant we set up for radiation 25 observation as right censored, 21 as interval censored and 20 as exact. Likewise, for R+C the set up to be 13 observation as right censored, 35 as interval censored and 20 as exact. The result of this data set will be analysis as interval censored and PIC as show in the next section.

Table 1: Time to cosmetic deterioration in breast cancer patients with two treatments

Radiotherapy
[0,8]; (0,5]; (4,11]; (5,12]; (5,11]; (6,10]; (7,16]; (7,14]; (11,15]; (11,18]; ≥15; ≥17; (17,25]; (17,25]; ≥18; (19,35]; (18,26]; ≥22; ≥24; ≥24; (25,37]; (26,40]; (27,34]; ≥32; ≥33; ≥34; (36,44]; (36,48]; ≥36; ≥36; (37,44]; ≥37; ≥37; ≥37; ≥38; ≥40; ≥45; ≥46; ≥46; ≥46;≥46; ≥46; ≥46; 4≥6; ≥46; 10; 23; 20; 37; 36; 20; 30; 20; 18; 30; 44; 23; 29; 15; 20; 22; 15; 45; 41; 38
Radiotherapy + Chemotherapy
(0,5]; (4,9]; (4,8]; (5,8]; (8,12]; (8,21]; (10,35];(10,17];(11,13];≥11;(11,17];≥11;(11,20];(12,20];≥13;(13,39];≥13;≥13;(14,17];(14,19];(15,22];(16,24];(16,20];(16,24];(16,60];(17,27];(17,23];(17,26];(18,25]; (18,24]; (19,32]; ≥21; (22,32]; ≥23; (24,31]; (24,30]; (30,34]; (30,36]; ≥31; ≥32; (32,40]; ≥34; ≥34; ≥35; (35,39]; (44,48]; ≥48;16; 25; 14; 12 ;24; 28; 26; 18; 40; 13; 21; 17; 27; 21; 22; 27; 9; 20; 40; 14

6. RESULTS

6.1 INTERVAL CENSORED (IC) DATA

We used the midpoint imputation to obtained estimated survival function for the two treatments R and R+C. Both estimates approach in Figure 1 show similar results compared with the one obtained by Turnbull. However, the patient in the R+C group develop breast retraction earlier than those in the R group, suggesting that our midpoint approach provides an acceptable approximation to the estimate. The left and right point imputation is used to obtain the estimated survival function for the two treatments R and R+C as shown in Figure 2. The Figure show that in the two treatments similar results are obtained by both left and right point as compared with the one obtained by Turnbull. Also, the patient in the R+C group develop breast retraction earlier than those in the R group, which indicate that our left and right point are better.

Similarly, the result obtained by using Random imputation mean imputation and median imputation in Figure 3 and 4, are the same as Turnbull. Moreover, the result obtained by mean and median imputation is significant with respect to the smallest value of P-value as shown in Table 2

Table 2: The P-value estimated based on Interval Censored (IC) Data

Methods	P-value
Turnbull	0.006384
Midpoint	0.004650
Left-point	0.010810
Right-point	0.037530
Random	0.008570
Mean	0.003329
Median	0.003285

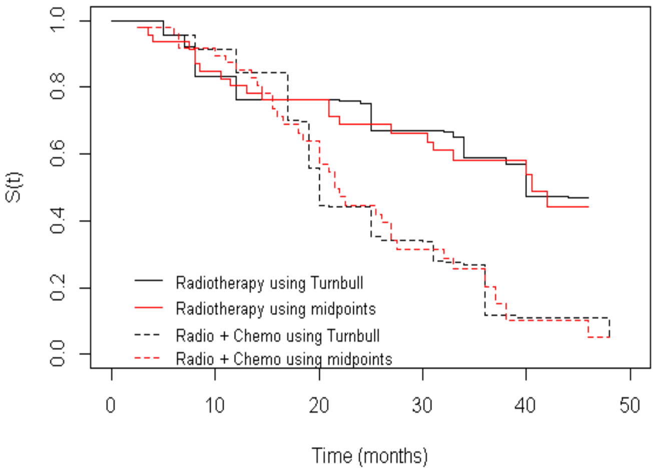

Figure 1: Estimated of Survival function obtained by Midpoint vs Turnbull Based on IC data.

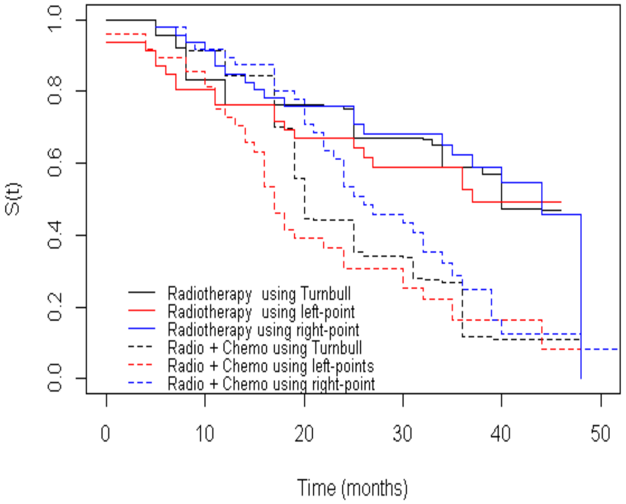

Figure 2: Estimated of Survival function obtained by Left & Right point Imputation vs Turnbull Based on IC data.

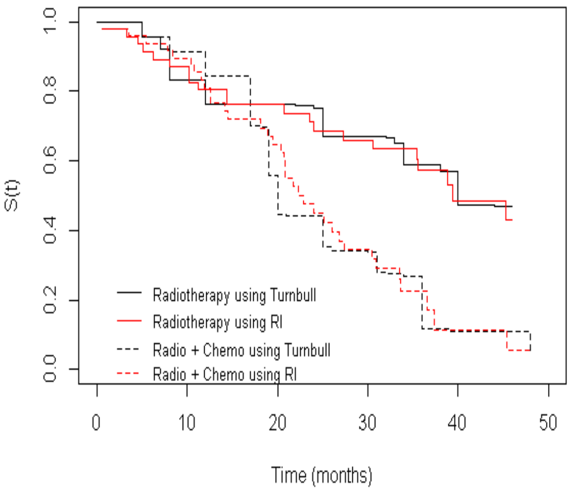

Figure 3: Estimated of Survival function obtained by Random Imputation (RI) vs Turnbull Based on IC data.

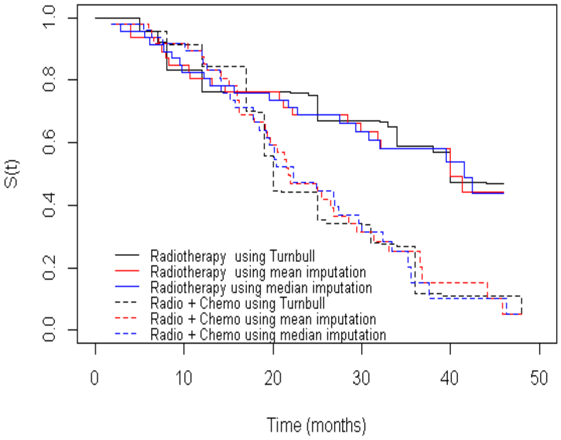

Figure 4: Estimated of Survival function obtained by Mean & Median Imputation vs Turnbull Based on IC data.

6.2 Partly-Interval Censored (PIC) data

For the partly interval censored data. Figure 5, 6, 7 and 8 show the results obtained by midpoint, left & right point, random and mean & median imputation, respectively. These results are almost similar to the one obtained by Turnbull methods as well as in Figure 1, 2, 3, 4. However, the random and mean & median imputation show better results compared to others type of imputations with respect to their P-value.

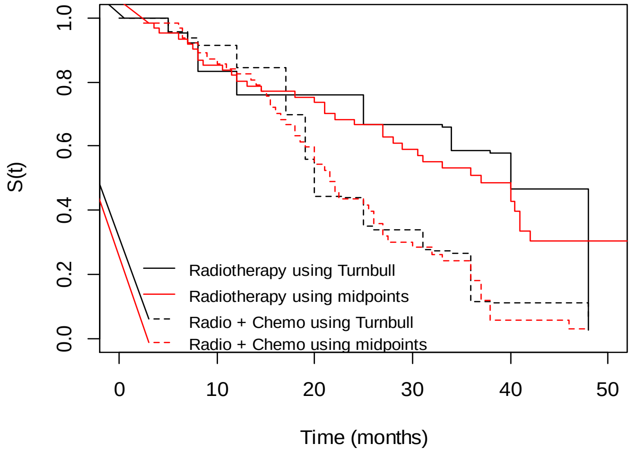

Figure 5: Estimated of Survival function obtained by Midpoint vs Turnbull Based on PIC data.

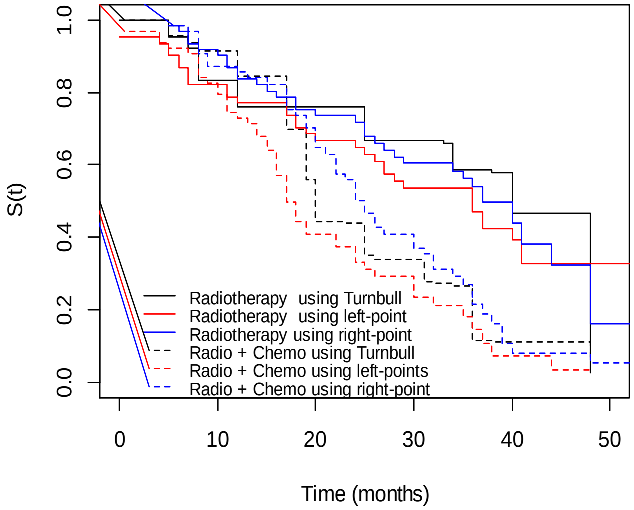

Figure 6: Estimated of Survival function obtained by Left & Right point Imputation vs Turnbull Based on PIC data.

Table 3: The P-value estimated based on PIC Data

Methods	**P-value**
Turnbull	0.001349
Midpoint	0.061240
Left-point	0.090400
Right-point	0.159500
Random	0.048200
Mean	0.003061
Median	0.003358

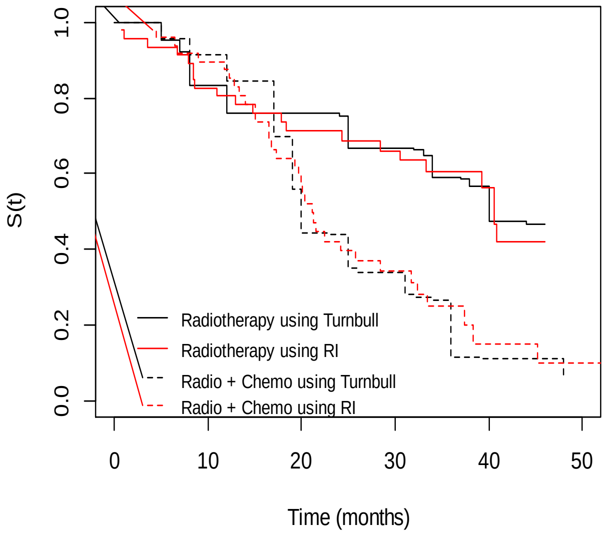

Figure 7: Estimated of Survival function obtained by Random Imputation (RI) vs Turnbull Based on PIC data.

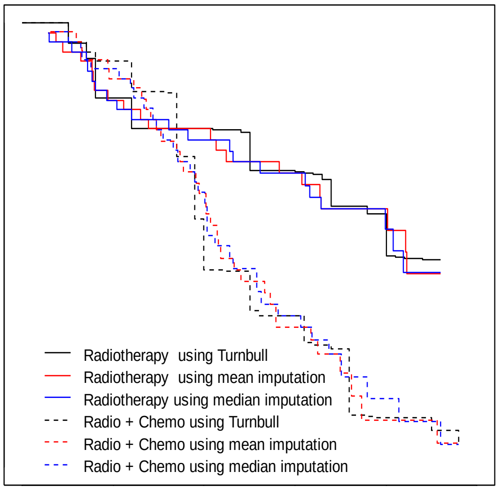

Figure 8: Estimated of Survival function obtained by Mean & Median Imputation vs Turnbull Based on PIC data.

7. CONCLUSION

We have proposed a simple modification of estimating survival function for partly-interval censored data using nonparametric estimate based on imputation techniques. Modification of breast cancer data is used and R software also used to obtain the results. Our results are fund to be similar to the one obtained by Turnbull. However, based on partly interval censored data, the random imputation and mean & median imputation show better results compared with others imputation techniques as well as Turnbull with respect to their the smallest P-value.

REFERENCES

[1] Alharpya,A.M.and Ibrahimb, N.A.(2013a). “Parametric tests for partly-interval censored failure time data weibull distribution via multiple imputation”, *Journal of Applied Science*, 13(4), pp 621-626.

[2] Alharpya,A.M.and Ibrahimb, N.A.(2013b). “A two sample parametric test for partly interval censored datawith non-proportional hazard”, *Mathematical Problems in Engineering*, Vol: 2014, Article ID 702847.

[3] Aragón,J.and Eberly, D.(1992).“On convergence of convex minorant algorithm for distribution estimation with interval-censored data”.J. Comput. Graph.Statist., 1: 129-140.

[4] Chen, L., and Sun, J. (2010) “A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model”. Computational Statistics & Data Analysis, 54(4): 1109-1116.

[5] Elfaki, F.A.M. (2012). “Parametric Cox’s Model for Partly Interval-Censored Data with Application to AIDS Studies,” *International Journal of Applied Physics and Mathematics*, 2(5), pp 352-354.

[6] Elfaki, F.A.M. Abobakar, A. Azram, M. et al. (2013). “Survival Model for Partly Interval-Censored Data with Application to Anti D in Rhesus D Negative Studies” *World Academy of Science, Engineering and Technology*, Vol: 77: 877-880.

[7] Finkelstein, D. M., and Wolfe, R. A. (1985). “A semi parametric model for regression analysis of interval-censored failure time data.” *Biometrics*, 41:933-945.

[8] Gauvreau, K. De Gruttola, V. Pagano, M. (1994). “The effect of covariates on the induction time of AIDS using improved imputation of exact seroconversion times.” *Statistics in Medicine*,13: pp 2021-2030.

[9] Geskus, R.B. (2001). “Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored”. *Statistics in Medicine*,20: pp 795-812.

[10] Goggins W. B, Finkelstein D. M, Zaslavsky A. M. (1999) “Applying the Cox proportional hazards model for analysis of latency data with interval censoring”. *Stat. Med.* 18 2737–2747**.**

[11] Groeneboom, P. (1991). Nonparametric maximum likelihood estimators for interval censoringand deconvolution.Technical Report 378, Department of Statistics, Stanford University.

[12] Groeneboom, P. and Wellner, J.A. (1992).Information Bounds and Nonparametric Maximum Likelihood Estimation, DMV Seminar Band 19, Birkh¨auser, Basel.

[13] Guure, C.B. and Ibrahim, N.A. (2013)“Generalized Bayesian non-informative prior estimation ofWeibull parameter with interval censoring”. *ScienceAsia*, 39S: 75–79.

[14] Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. Statistica Sinia, 9: 501 – 519.

[15] Jongbloed, G. (1995). “Three statistical inverse problems”.Ph.D. thesis, Delft Technological University, The Netherlands.

[16] Jongbloed, G.(1998). “The iterative convex minorant algorithm for nonparametric estimation”.*J. Comput. Graph. Statist.*,7:310-321.

[17] Kim, J. S. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. Journal of the Royal Statistical Society, 65: 489- 502.

[18] Kumar, P. (2010) Modeling and analysis of degradation data. Indian Institute of TechnologyBombay,Available:http://www.academia.edu/266427/Modelling_and_Analysis_of_Degradation_Data [Accessed: 5/1/2013].

[19] Law, C.G. Brookmeyer, R. (1992). “Effects of mid-point imputation on the analysis of doubly censored data”.*Statistics in Medicine*, 11: pp 1569-1578.

[20] Leffondre K., Touraine, C., Helmer, C., and Joly, P. (2013) “Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model?”. International Journal of Epodiomology, Vol. 42, 2013, 1177-1186.

[21] Liu, K.J. Darrow, W.W. and Rutherford,G.W.(1988) “A model-based estimate of the mean incubation period for AIDS in homosexual men”. *Science*, 240:1333-1335.

[22] Mariotto, A.B. Mariotti, S. Pezzotti, S. (1992).“Estimation of the acquired immunodeffciency syndrome incubation period in intravenous drug users: A comparison with male homosexuals. *American Journal of Epidemiology*, 135: pp 428-437.

[23] Odell, P. M., Anderson, K. M. and D’Agostino, R. B. (1992). “Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model.” *Biometrics*,48:951-959.

[24] Pan, W. (2000). “A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies”. *Biometrics*,57:1245-1250.

[25] Peto R. and Peto, J. (1972).“Asymptotically Efficient Rank Invariant Test Procedures”.*J R. Statist. Soc*., Series A135: 187-220.

[26] Schick, A. and Yu, Q. (2000)“Consistency of the GMLE with mixed case interval-censored data”. *Scandinavian Journal of Statistics*, 27: 45-55.

[27] Sen,B. and Banerjee,M. (2007). “A pseudolikelihood method for analyzing interval censoreddata”. *Biometrika*, 94:71-86.

[28] Singh, R.S.and Totawattage, D.P. (2013) “The Statistical Analysis of Interval-Censored Failure Time Data with Applications”.*Open Journal of Statistics*, 3: 155-166.

[29] Song, S. (2004). “Estimation with univariate “mixed case” interval censored data”. *Statist.Sinica*,14:269-282.

[30] Sun, J., Zhao, Q., & Zhao, X. (2005). “Generalized log-rank test for interval censored failure time data.” Scand. J. Stat. 32: 49-57.

[31] Tillmann, H.L. Heiken, H.Knapir-Botor, A. (2001). “Infection with GB virus C andreduced mortality among HIV-infected patients”.*New England Journal of Medicine*, 345: 715-724.

[32] Turnbull, B. W. (1976). “the empirical distribution function with arbitrary grouped censored and truncated data”*J R. Statist. Soc.*, 38: 290-295.

[33] Yu, Q. Li, L. and Wong, G. Y. C.(2000). “On consistency of the self-consistent estimator of survival functions with interval-censored data”.*Scandinavian Journal of Statistics*, 27: 35-44.

[34] Vander Vaart, A.W.and Wellner, J.A. (2000).“Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes”.*High Dimensional Probability II* Vol: 47: 115-133.

[35] Xiang, J. Wünschmann, S. Diekema, D.j. (2001).“Effect of coinfection with GB virus C on survival among patients with HIVinfection”.*New England Journal of Medicine*,345: pp 707-714.

[36] Zhang, W., Zhang, Y., Chaloner B. K, and Stapletonc, J. T. (2009) “Imputation methods for doubly censored HIV data”, J Stat Comput Simul. 79(10): 1245–1257.

[37] Zhao, X. Zhao, Q.J. Sun, (2008). “Generralized log-rank test for partly interval-censored failure time data,” *Biometrical Journal*, Vol. 3:375-385.