World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:5, No:3, 2011

Parametric and Nonparametric Analysis of Breast Cancer Treatments

Chunling Cong, Chris.P.Tsokos

***Abstract*—**The objective of the present research manuscript is to perform parametric, nonparametric, and decision tree analysis to evaluate two treatments that are being used for breast cancer patients. Our study is based on utilizing real data which was initially used in “Tamoxifen with or without breast irradiation in women of 50 years of age or older with early breast cancer” [1], and the data is supplied to us by N.A. Ibrahim “Decision tree for competing risks survival probability in breast cancer study” [2]. We agree upon certain aspects of our findings with the published results. However, in this manuscript, we focus on relapse time of breast cancer patients instead of survival time and parametric analysis instead of semi-parametric decision tree analysis is applied to provide more precise recommendations of effectiveness of the two treatments with respect to reoccurrence of breast cancer.

***Keywords*—**decision tree, breast cancer treatments, parametric analysis, non-parametric analysis

I. Introduction

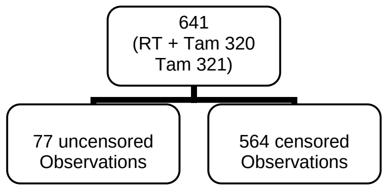
EXTENSIVE literature and studies can be found related to whether radiation shows a benefit to breast cancer patients with respect to relapse time. It is clear that radiation makes a difference in recurrence for some women. However, the potential side effect of heart damage from breast radiation makes it desirable to avoid radiotherapy unless it is absolutely necessary. Therefore, it is of great importance to identify the patients who could potentially benefit from radiation and those who would be put at higher risk for receiving radiation treatment. The aim of the present research is to perform parametric, nonparametric, and decision tree analysis to answer the above question. Our parametric and nonparametric analysis confirms the overall advantage of combined radiation and tamoxifen (RT+Tam) over tamoxifen (Tam) alone in reducing the probability of relapse; however, after utilizing decision tree analysis in conjunction with survival analysis of relapse time of breast cancer patients, we have concluded under some conditions, giving both treatments to patients without considering the clinicopathological characteristics could be negatively effective or catastrophic.

Chunling Cong is a doctoral student in the Department of Mathematics and Statistics at University of South Florida, Tampa, FL 33613 USA (e-mail: ccong@mail.usf.edu).

Chris P. Tsokos, is with University of South Florida. He is a distinguished university professor, president of IFNA, and director of the statistics graduate program, Tampa, FL 33613 USA (e-mail: profcpt@cas.usf.edu).

II. Data

Between December 1992 and June 2000, a total of 769 women were enrolled and randomized, of which 386 received combined radiation and tamoxifen (RT+Tam), and the rest, 383, received tamoxifen-alone (Tam). The last follow up was conducted in the summer of 2002. Only those 641 patients enrolled at the Princess Margaret Hospital are included: 320 and 321 in RT+Tam and Tam treatment group, respectively.



Patient treatment data

This censored data consists of 77 uncensored observations and 564 censored observations. The censored observations are mostly due to two reasons: (1) the breast cancer patient emigrated out of the study area; (2) the individual survived (did not experience occurrence) past the end of the study period. Due to the fact that nearly 90% of the data are censored observations, we take into consideration two datasets, 77 uncensored dataset, and 641 censored dataset for later analysis.

In the original data, three relapse events are recorded: local relapse, axillary relapse and distant relapse. The original dataset was used to analyze competing risks (also called multiple causes of death) including relapse, second malignancy, and other causes of death. Since in the present study we are interested in the relapse time regardless of the reoccurrence type, minimum time of the three types of relapse is chosen for analysis purpose, and the values of censoring indicator variable are adjusted accordingly. Variables assessed at the time of randomization are: **pathsize**(size of tumor in cm) ; **hist**(Histology: DUC=Ductal, LOB=Lobular, MED= Medullar, MIX=Mixed, OTH=Other); **hrlevel**(Hormone receptor level: NEG=Negative, POS=Positive); **hgb**(Hemoglobin g/l); **nodediss**(Whether axillary node dissection was done: Y=Yes, N=No); **age**(Age at diagnosis in years). All these attributable variables will be used in the modeling of breast cancer in a separate study where various statistical models are used to identify the significant prognostic factors in the relapse of breast cancer, as well as the interactions between the variables and ranking of significant individual attributable variables and interactions.

III. Nonparametric Analysis

Kaplan-Meier estimator [3] (also known as the product limit estimator) estimates the survival function from survival related data. In many medical researches, it is used to measure the portion of patients living for a certain amount of time after treatment. Kaplan-Meier is useful when we have censored data, and it is equivalent to the empirical distribution when no truncation or censoring occurs.

Let $S(t)$ be the probability that an individual will not have reoccurrence of breast cancer after time t. For a sample of size n, denote the observed times until death of n sample members as $t_1 \le t_2 \le t_3 \le \ldots \le t_n$. Then the nonparametric Kaplan-Meier estimator of the survival function is estimated by :

$$\hat{s}(t) = \prod_{t_i \le t} \frac{n_i - d_i}{n_i} \qquad (1)$$

where n_i is the number of survivors just prior to time t_i, and d_i is the number of deaths at time t_i.

Kaplan-Meier estimates of the survival curves of relapse time for the two treatment groups are shown in Fig. 1.

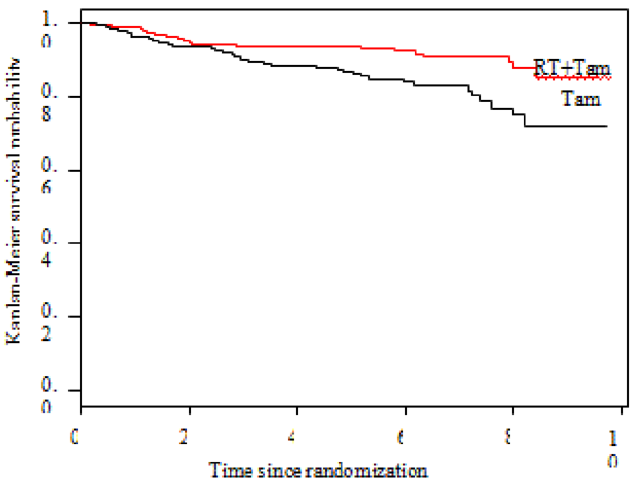


Fig. 1 Survival curves of two treatment groups

Kaplan-Meier is a nonparametric procedure for estimating the survival curve; however, it is not commonly used to compare the true mean effectiveness of the two treatments. In the present study, we perform actual nonparametric analysis utilizing Wilcoxon rank sum test and Peto & Peto modification of the Gehan-Wilcoxon test. We proceed in nonparametric direction for comparison purpose with the results obtained using parametric analysis. Utilizing the two different nonparametric tests, we found the information in Table 1, which shows that the combination of the two treatments (RT+Tam) is more effective than using the single treatment (Tam) which is consistent with Fig. 1.

TABLE I
Test The Difference Of Meas Of Two Treatments

	Chi-Square	Degree of freedom	P-value
Log-rank	9.8	1	0.0017
Peto & Peto	9.6	1	0.00197

IV. Parametric Analysis

First, censored dataset which consists of 641 patients are analyzed, and the characteristic of the behavior of relapse time in RT+Tam arm is investigated through goodness of fit tests. The best probability distribution is the lognormal distribution, with corresponding maximum likelihood estimator (MLE) of the following form $\hat{\mu}$ =5.148, $\hat{\sigma}$ =2.47 (as shown in Table 2). A graphical presentation of the cumulative distribution function (CDF) is shown by Fig. 2 where Kaplan-Meier curve and its 95% confidence band, as well as CDF of the fitted lognormal distribution are plotted.

TABLE II
Estimated Parameters And Log-Likelihood Of Lognormal Distribution

	$\hat{\mu}$	$\hat{\sigma}$	Log-likelihood
Totality	4.101	2.04	-367
RT+Tam	5.148	2.47	-134.4
Tam	3.491	1.79	-227.3

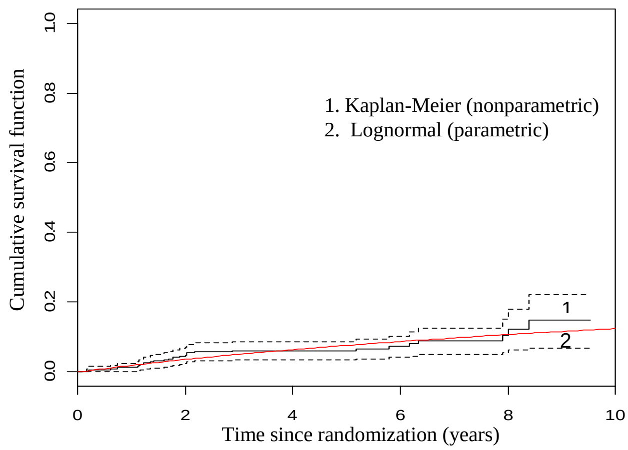


Fig. 2 Fitted lognormal CDF curve for RT+Tam

As can be seen from the above graph, lognormal probability distribution seems to be a good fit for the relapse time of breast cancer patients in RT+Tam, and the survival curve from the lognormal probability distribution with estimated parameters is very close to the Kaplan-Meier survival curve and it is within the 95% confidence band constructed from Kaplan-Meier survival curve.

Similarly, we perform a parametric analysis for patients in Tam arm. It has been proven through goodness-of-fit test that the subject data follows a lognormal distribution as well, with MLE of $\hat{\mu}$ =3.419, $\hat{\sigma}$ =1.79 (as shown in Table 2). Therefore, the final estimated form of the lognormal probability distribution is given in Table 2 and a graphical form of the cumulative distribution function is given in Fig.3.

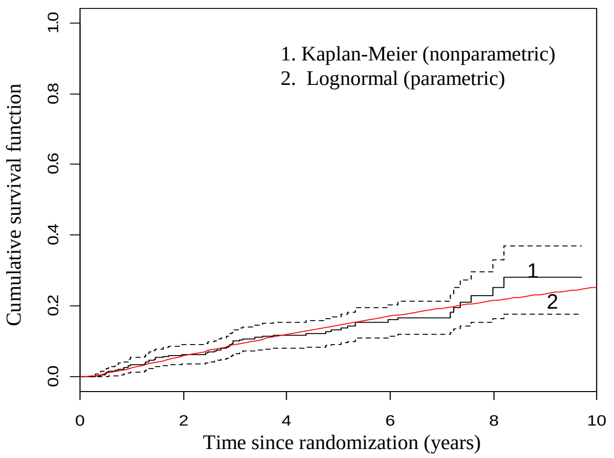


Fig. 3 Fitted lognormal survival curve for Tam

Since relapse time in RT+Tam and Tam arm both follow lognormal probability distribution, the log-likelihood ratio test can be applied to test hypothesis

$$H_o : \mu_1 = \mu_2 = \mu \text{ vs. } H_1 : \mu_1 \neq \mu_2$$

The log-likelihood ratio test statistic is given by

$$T = -2[l(\mu,\mu) - l(\mu_1 - \mu_2)] = 10.6$$

with one degree of freedom, and from the Chi-square distribution table, p-value is between 0.05 and 0.001. Thus, there is significant difference between the locations of the two treatment groups, which is consistent with the conclusion using nonparametric tests.

While for the uncensored dataset of the 77 breast cancer patients, of which 26 are treated with RT+Tam and 51 with Tam alone, in order to perform goodness of fit test to identify the PDF of the 26 patients, we employ bootstrapping technique to increase the sample size of the RT+Tam arm. Through goodness of fit tests including Kolmogorov-Smirnov, Anderson-Darling and Chi-Square tests, the best fit for RT+Tam is log-logistic probability distribution while the best for Tam arm is general Pareto probability distribution. Considering the difference in probability distributions of the two groups, further analysis or tests are not conducted to check the mean difference in relapse time. Since consistent results were obtained using nonparametric and parametric tests with regard to the censored dataset, we only considered the censored dataset for the subsequent analysis. However, as we will see in the following discussion, after applying decision tree analysis to partition the subject data as a function of the tumor size, age of patient and haemoglobin, the findings of the two treatments give contradictory results which could be quite misleading in the treatment of breast cancer patients as the nonparametric and parametric analysis indicates.

V. Decision Tree Analysis

The clinicopathological characters of breast cancer patients are heterogeneous. Consequently, the survival times are different in subgroups of patients. Decision tree analysis [4]-[9] is used to homogenize the data by separating the data into different subgroups on the basis of similarity of their response to treatment. The general goal of such applications is to identify prognostic factors that are predictive of survival outcome and time to an event of interest (relapse time in this study). For example, a tree-based decision analysis enables the natural identification of prognostic groups among patients, using information available regarding several clinicopathologic variables. Such groupings are important because patients treated with RT+Tam and Tam present considerable heterogeneity in terms of relapse time, and the groupings allow physicians to make early yet prudent decisions regarding adjuvant combination therapies.

The concept of exponential decision tree analysis [10] is to reduce the impurity within nodes by splitting based on covariates using a specified loss function. Assuming the hazard rate within a given node is constant, $h(y) = \lambda_j$ for all y in group j, and then the survival function within each node is an exponential function. The split point is selected so that the loss among the possible binary splits defined by the covariates are minimized. The loss function for a node t is given by

$$R(t) = -\hat{L}(t) = D_t - D_t \log(D_t / Y_T) \tag{2}$$

where $D_t = \sum_i d_i$ is the number of complete observations at the node and $Y_t = \sum_i y_i$ is the total observed time.

Considering our main focus here is to compare the two treatments instead of analyzing each treatment alone, the maximum tree depth is set to be 3 with complexity parameter 0.02. The trees of RT+Tam and Tam are shown in Fig.4 and Fig.5 as follows:

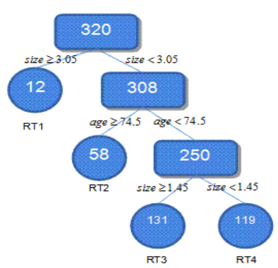


Fig. 4 Radiation +Tamoxifen

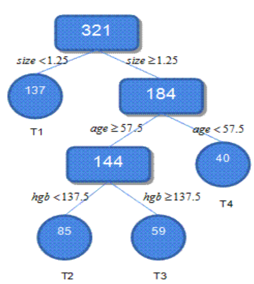


Fig. 5 Tamoxifen

RT+Tam arm is divided into 4 groups denoted by RT1,RT2,RT3,RT4 from the left to the right; Tam arm is divided into 4 groups denoted by T1,T2,T3,T4 from the left to the right. To further investigate the survival curves of a subgroup from different treatment arms, Kaplan-Meier survival curves are plotted in Fig. 6.

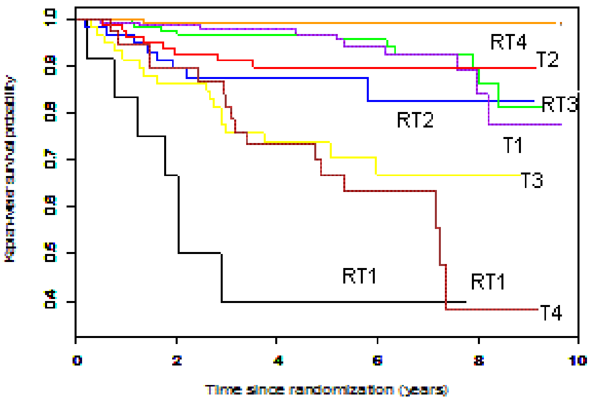


Fig. 6 Survival time for different subgroups

Using decision tree analysis we conclude that giving radiation to a patient whose tumor size exceeds 3.05cm would be catastrophic as has been shown in Fig.6 since patients in RT1 are most likely to relapse. Furthermore, treatment Tam is more effective than treatment RT+Tam with respect to relapse time has also been shown by the survival curves of T2 and RT2. In addition, we can conclude that by using decision tree analysis and the corresponding survival analysis, we can group the breast cancer patients into three clusterings that identify the effectiveness of treatment RT+Tam versus treatment Tam. For example, the survival curve of RT3 is very close to that of T1, which suggests that for patients whose age is under 74.5 and have tumor size between 1.45cm and 3.05 cm, RT+Tam shows no advantage over Tam. Thus, it would be desirable for this patient not to consider receiving radiation.

We summarize below when RT+Tam and Tam are almost equally effective

(1)RT4, T2, RT3, RT2, T1

(2)T3, T4

(3)RT1

Thus, our findings are important in guiding the physicians to recommend tamoxifen alone without radiation rather than a combined treatment of tamoxifen and radiation when they are equally effective to breast cancer patients with certain size of tumor, age and hemoglobin level.

VI. Conclusion

Although overall parametric and nonparametric comparisons of RT+Tam and Tam arms show that the combination of radiation and tamoxifen is more effective than tamoxifen alone with regard to the relapse time of a breast cancer patient, a decision tree analysis for censored data reveals that the heterogeneity of clinicopathological characteristics lead to important difference between subgroups of the two treatment groups, thus affecting the decision making process in choosing suitable treatment for breast cancer patients.

Acknowledgement

We wish to thank N.A Ibrahim for supplying us the source of the data that made the subject study possible. We also wish to express appreciation to Dr. James Kepner, Vice President of the American Cancer Society for our useful discussion on the present study.

References

[1] A. W. Fyles, D.R. McCready, et al., “Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer”, New England Journal of Medicine 351, pp. 963-970, 2004.

[2] N.A Ibrahim, et al “Decision tree for competing risks survival probability in breast cancer study”, International Journal of Biomedical Sciences, Volume 3 Number 1, 2008.

[3] Kaplan, E.L.; Meier, Paul. “Nonparametric estimation from incomplete observations”. J. Am. Stat. Assoc. 53, 457-481, 1958.

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and Regression Trees”, New York; Chapman and Hall, 1984.

[5] J. R. Quinlan, “C4.5: Program for Machine Learning”, California: Morgan Kaufmann, 1992.

[6] M. R. Segal,”Regression trees for censored data”, Biometrics 44, pp.35-47, 1988.

[7] X.G. Su and J.J. Fan, “Multivariate survival trees: a maximum likelihood approach based on frailty models”, Biometrics 60, pp. 93-99, 2004.

[8] F. Gao, A. K. Manatunga, and S. Chen, “Identification of prognostic factors with multivariate survival data”, Computational Statistics and Data Analysis 45, pp. 813-824, 2004.

[9] LeBlanc, M., Crowley, J., “Survival trees by goodness of split”. Journal of the American Statistical Association 88, 457–467, 1993.

[10] R. Davis and J. Anderson,”Exponential survival trees”, Statistics in Medicine 8, pp 947-962, 1989.

Chunling Cong is currently a doctoral student in Mathematics and Statistics at University of South Florida. She is also a teaching assistant in the department.

Chris. P. Tsokos was born near Kalavrita, Greece, and emigrated to America in his teens. He attended the Universities of Rhode Island and Connecticut, MIT, and Penn State. He has a doctorate in statistics and has taught at University of Rhode Island, Virginia Polytechnic Institute, and the University of South Florida, where he is currently a Distinguish University Professor of Mathematics and Statistics and has been dissertation advisor for more than 30 doctoral students and advisor for more than 75 master's candidates. He has authored over 275 research and 35 technical articles, fifteen textbooks and monographs, and four chapters in books of research.

Dr. Tsokos has earned grants of over a million and a half dollars from the National Science Foundation, NASA, National Institute of Health, the Office of Naval Research, the U.S. Air Force, and the U.S. Army. He has also served as advisor and consultant to many government agencies as well as several Fortune 500 companies. He has served as officer of numerous professional organizations, co-edited three academic journals and monograph series, and served as associate editor of more than ten other journals. He Co-founded and continues to shepherd the Urban Scholars program, which helps provide underprivileged youth with academic coaching and technology resources. He has been named a Distinguished Scholar at USF and has been recognized many times for his academic and humanitarian work. For the past thirteen years he has been serving as founding president of the American Foundation of Greek Language and Culture, AFGLC. www.afglc.org