Available online at http://www.journalijdr.com

International Journal of
DEVELOPMENT RESEARCH

IJDR

ISSN: 2230-9926

International Journal of Development Research
Vol. 06, Issue, 01, pp. 6439-6443, January, 2016

Full Length Review Article

A STUDY ON SURVIVAL ANALYSIS

*Sujatha, V. and Kalpanapriya, D.

Department of Mathematics, VIT University, Vellore, India

ARTICLE INFO

Article History:
Received 28th October, 2015
Received in revised form
14th November, 2015
Accepted 29th December, 2015
Published online 31st January, 2016

Key Words:
Survival analysis,
Review,
Statistics.

ABSTRACT

Survival times are used to compare the treatments. Survival data, a special type of data, have to be analyzed with special methods. when survival times are analyzed without the use of special techniques, or when the underlying assumptions were not taken into an account, then faulty interpretation may result. Readers should know these pit-falls and be able to judge for themselves whether the chosen analytical method is correct. The present article is based on textbooks of statistics, a selective review of the literature, and the authors own experience. This article suggest the guidelines for the presentation of survival analyses in medical journals. These would be the complement on statistical guidelines recommended by several medical journals.

INTRODUCTION

Survival analysis has found widespread applications in medicine in the last 10-15 years. The literature on the Survival Models and Comparison of Survival Models for estimating the survival function is growing day-by-day. Extensive research works in this area were carried out by many researchers. when careful examinations were made on the use of statistical methods in survival analysis, there has not been any published review of the use of survival analysis methods in medical journals. A systematic review gives the appropriateness of the application and presentation of survival analyses in medical journals, and presentation is critically relevant for much of the clinical diabetic literature.

Univariate Analyses

Models fitted to survival data may involve parametric or non-parametric forms for the hazard function. This depends on whether this form is defined (up to a small number of unknown parameters. In univariate survival analysis, the length of survival was examined in relation to only one explanatory variable at a time, hence ignoring the simultaneous effects of other variables. It compares the survival in two or more groups, one of the most familiar method is the log rank test, which also goes under several other names including Mantel, Mantel- Haenszel, generalised Savage and Mantel-Cox.

There are also a class of tests (referred to here as weighted log rank) which allow events occurring at different times to have differing weights in the computation (Harrington and Fleming, 1982), the best-known names being the generalised Wilcoxon and the Gehan. Whenever the variable examines three or more ordered categories, the more appropriate log rank test is a trend (which seeks monotonic relationship instead of just heterogeneity). Cox proportional hazards regression model (Cox, 1972) with a single explanatory variable is an alternate in the place of or in an addition to that of log rank test. The reporting observed events which are calculated by ignoring the differing lengths of follow-up are kept as simple indicators which are easily misinterpreted.

Multivariate Analyses

when survival analysis as multivariate, the survival probability was examined in relation to at least two explanatory variables simultaneously, however in Cox regression analysis with baseline covariates (time-fixed), Cox regression analysis with covariates measured over time (time-dependent), the fitting of a Weibull model, multivariate logistic regression, adjusted Kaplan-Meier or stratified log rank analyses, most of the authors used the stepwise analysis to model their assumptions and the chi-square test for goodness of fit for their final model. The papers on which reports the results of multivariate analyses, among them half of which included estimates of some sort and it gave a standard errors or confidence intervals. Most of the estimates were provided by many authors were univariate analysis, in which only few provided an adequate

*Corresponding author: **Sujatha***
Department of Mathematics, VIT University, Vellore.India

summary. mostly it was unclear and they didn't show how exactly the multivariate analyses had been carried out.

Parametric semi parametric and non parametric models

Parametric survival model makes assumptions about the functional form of the probability distribution and the way that the explanatory variables influence the risk of ratifying. The first assumption deals with the functional form of the probability distribution. The probability distribution summarizes how the probability of ratifying changes over time. One way to represent the probability distribution is the hazard function. The hazard function can be thought of as the instantaneous probability of ratifying, conditional on not having ratified so for. When the functional form of the distribution is chosen then it imposes constraint on the shapes the distribution, but not fixing it completely. For instance, the simplest functional form of the probability distribution is to assume that the hazard is constant over time. This would mean that risk is always the same.

A simple assumption is the proportional hazard assumption, which are all used in the Exponential, the Weibull and the Gompertz models. The exponential distribution was studied first in connection with the kinetic theory of gases (Clausius, 1858). It plays a pivotal role in the theory of point processes (Cox and Isham, 1980; Cox and Lewis, 1966). The Weibull distribution was introduced by Fisher and Tippett (1928) in connection with extreme value distributions; Weibull (1939a, b) studied it in an investigation of the strength of materials. Wiley.Nassar and Eissa (2003, 2004), studied a two-parameter Exponenciated weibull (EW) model of the form They gave some of its properties and estimated the parameters by using the maximum likelihood and Bayes methods based on type II censored data. They used the squared-error and linear exponential loss functions and an informative prior to obtain the Bayes estimates. The two parameter model, Exponenciated Exponential Model (EEM) is defined as a particular case of Gompertz-verhuslt distribution function (Ahuja and Nash,1967). The EEM has been discussed by (Gupta and Kundu, 1999).The cumulative distribution function of EEM (Gupta and Kundu, 1999 to 2003 is defined by $F(t, \alpha, \lambda) = (1- e^{-\lambda t}) ^{\alpha}$. The EE density varies significantly depending on the shape of the parameter with $\lambda = 1$.

In a series of papers, R.D. Gupta and D. Kundu (1999, 2001a, 2001b, 2002, 2003a, 2003b, 2004), R.D. Gupta et (2002), Kundu and R.D. Gupta (2005) and Kundu *et al* (2005) concentrated on the study of the exponentiated exponential (EE) or what they also called generalized exponential distribution where G is the exponential CDF. In one of their papers, Kundu and R.D. Gupta (2005) stated that “the two-parameter (exponenciated exponential) distribution is an alternative to the well-known two-parameter gamma, two parameter Weibull or two-parameter lognormal distributions”. G. S. Mudholkar and D. K. Srivastava, explained the “Exponentiated Weibull family for analyzing bathtub failure-rate data. Most of the papers reveals the parametric survival models are statistically more powerful than nonparametric or semi parametric models. Hazard rate is the probability of an individual survives at t experienced the target event at specified period greater than t. The shape of the hazard rate changes with respect to time and It varies from one situation to the another situation. Parametric models are assuming some underlying shape to the survival curve.

Each parametric model specifies a particular shape for the hazard rate i.e. the time dependency. The exponential model assumes a flat hazard; the Weibull assumes a monotonic hazard; the Lognormal and Log logistic assume a non-monotonic hazard. If the characterization of the underlying time-dependency is accurate then a suitable distribution function may be selected. Parameter estimates will generally more precise than estimates from semi parametric and nonparametric models where the underlying time-dependency is left unspecified. In most of the papers, the shape of the hazard function when there are no covariates it may not be a good guide to the shape of the hazard function when there are covariates, parametric models differ not only in terms of the assumptions made about the shape of the hazard rate but also in terms of their specifications and interpretations. The product-limit estimate of the survival function has been in use since the early 1900s. The expression for the standard error of the Kaplan-Meier estimate was first given by Greenwood (1926). Kaplan and Meier (1958) method provides very useful estimation of survival probabilities and graphical presentation of survival distribution that help us to compare two or more survival distributions. Gehan (1965) has written classic reports on life-table analysis. The variance of Nelson-Aalen (1972) estimator was estimated by Aalen (1978) using counting process techniques.

Peterson (1977) expressed finite sample censored survivorship function as an explicit function of two empirical sub survival functions which has got strong consistency property. The Kaplan-Meier estimator for the survival function in the censored data problem can be expressed for finite samples as an explicit function of two empirical sub survival functions. Jan *et al., (*2005) attaches’ non censored rate as weights for censored observations, in case of high proportion of censoring and which makes the survival estimates less biased. Shafiq *et al.,* (2007) proposed a new weight that gives non-zero weight to the last censored observation, in order to avoid zero probability for the same. Many of the popular nonparametric two-sample test statistics for censored survival data, such as the log-rank (Mantel, 1966), generalized Wilcoxon (Gehan, 1965), and Peto-Peto (1972) test statistics, have been shown to be special cases of a general two-sample statistic, differing only in the choice of weight function (Tarone and Ware, 1977; Gill, 1980). This work has been extended to a general s-sample statistic (Tarone and Ware, 1977; Andersen *et al*., 1982) which includes the s-sample log-rank (Breslow, 1970) and generalized Wilcoxon (Prentice, 1978).

The Proportional Hazards (PH) model was first proposed by Cox (1972), who emphasized the log linear form for the multiplier. He derived the likelihood as a product of conditional probabilities. Estimators for the hazard functions were suggested by Cox (1975). Vaupel *et al.,* (1979) discussed the impact of heterogeneity in individual frailty on the dynamics of mortality. Kalbfleisch and Prentice (1980) pointed out the explicit interpretation in terms of the marginal likelihood of ranks. Schoenfeld (1982) suggested that the residuals can be plotted against time to test the proportional hazards assumption. Histograms of these residuals can be used to examine the fit and detect outlying covariate values. Lawless (1982) presents and illustrates statistical methods for modeling and analyzing life time data. Gill (1984) discussed how martingale techniques can be used to extend Cox's regression model and to derive its large sample properties. Lee (1992) suggested that if less than 50% of the observations are uncensored and the largest observation is censored, the median survival time cannot be estimated. Willett and Singer (1993) showed how discrete-time survival analysis can address questions about onset, cessation, relapse, and recovery. Hougaard (2000) presents four approaches to handle multivariate survival data and clarified the concepts both for simple survival data and multiple survival data. Ibrahim *et al.,* (2001) examined Bayesian approaches to survival analysis and also discussed several types of models, including parametric and semiparametric, proportional and non-proportional hazards, frailty and models with time-varying covariates.

Nardi and Schemper (2003) investigated the comparative performance of Cox and Parametric survival models under the typical condition of clinical studies. Pourhoseingholi *et al.,* (2007) compared Cox regression and Parametric models for survival of Patients with Gastric Carcinoma and concluded that in multivariate analysis, Cox and Exponential models behave similar. In univariate analysis he suggested in certain cases, the Lognormal regression behaves better among parametric models and it can lead to more precise results as an alternative to Cox model. Jiezhi Qi (2009) compared Proportional Hazards (PH) and Accelerated Failure Time (AFT) models and suggested that Cox PH model may not be the optimum approach. Using Maximum likelihood method and Akaike Information Criterion (AIC), Nakhee and Law (2009) found Weibull model to be the best parametric model fitted to people with a diagnosis of Human Immune Deficiency Virus (HIV) positive cases without Acquired Immune Deficiency Syndrome (AIDS). Ponnuraja and Venkatesan (2010) suggested that PH model is not always appropriate and that the AFT model in many applications provides a more appropriate modeling framework and have the added advantage of being straightforward to interpret than the PH model. Hayat *et al.,* (2010) compared the results of the survival analysis of the patients with breast cancer using Weibull, Gamma, Gompertz, Loglogistic and Lognormal Parametric models.

They observed the AIC values for the five distributions were very close to each other. The Gompertz distribution, which had the lowest AIC value, was determined as the most suitable method. Ramadurai and Ponnuraja (2011) proposed the Schoenfeld residual check plays a dominant role in validating the diagnostic check on the Cox PH model. Venkatesan *et al.,* (2011), identified risk factors and prognostic factors for breast cancer survival for patients treated under adjuvant and neo-adjuvant therapy. Grzenda (2012) indicates the main factors influencing unemployment duration using Bayesian Exponential Survival model. Therneau *et al.,* (2013), used Cox model in its ability to encompass covariates that change over time and Vallinayagam *et al.,* (2014) observed lognormal model is the suitable model for Breast Cancer Survival Data.

Regression and neural networks

Regression is a statistical technique to estimate the relationship between a dependent variable and two or more independent variable. The general regression model does not require an iterative training procedure. Regression coefficients are estimated by minimizing the sum of residuals. The standard error of the regression is based on the sum of the residuals. Regression is the attempt to explain the variation in a dependent variable using the variation in independent variable. There is substantial literature regarding the efforts made in the field of nonlinear regression. To measure the nonlinearity also to right of entry the adequacy of regression models with their estimation, certain work is available (see e.g., Beale, 1960; Guttman and Meeter, 1965). Bates and Wates (1980) presented measure of nonlinearity about the geometric behavior of the curvature. They have found two components of nonlinearity i.e., intrinsic nonlinearity (IN) and parameter effect (PE) nonlinearity. Bates and Wates (1980) examined the work of Beale (1960) and Box (1971) and showed that Beale's measure generally tend to underestimate the true nonlinearity, but the bias measure of Box is closely related to the parameter effect nonlinearity.

Neural Network is an tool to predict the diabetes of a patient. Data mining also popularly known as knowledge discovery to find the interrelation pattern among the data. It provides an useful information from large set of data bases. Neural networks are used for prediction with various levels of success. The advantage includes the automatic learning of dependencies only from measured data without any need to add further information. The predictive accuracy on a neural network is more than the Regression technique of human experts Neural networks can learn the dependency valid for a certain period. The knowledge stored in the form of Neural network are strongly non-linear dependent and even there is significant noise in the training set. The knowledge is not a comprehensible.

Smith *et al.* used the PID data set to evaluate the perceptron-like Adaptive learning routine (ADAP). This study had 576 cases in the training set and 192 cases in the test set. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances. Upadhyaya S., Farahmand K., Baker-Demaray F.,they Compare the Neural Network and Linear Regression. They classify the context of screening native American elders with diabetes. Kamer Kayaer, Tulay Yildirim 'Medical Diagnosis on Pima Indian Diabetes using general Regression Neural Networks'. They compare the linear General Regression Neural Network (GRNN) model with two different neural network structures, which are multilayer perceptron (MLP), radial basis function (RBF). They applied three tools to the Pima Indians Diabetes (PID), medical data The best result achieved on the test data is the one using the GRNN structure (80.21%). This is very close to one with the highest true classification result that was achieved by using the more complex structured ARTMAP-IC network (81%).Therefore they show that, general regression neural network (GRNN) can be a good and practical choice to classify a medical data. Similarly, Jankowski and Kadirkamanathan developed a radial basis function network suite called Inc Net which used 100 neurons and trained for 5,000 iterations. This approach yielded 77.6% accuracy.

Au and Chan in attempted to improve the correct classification percentage on the PID dataset by using a fuzzy approach. Au and Chan first represented the revealed regularities and exceptions using linguistic terms, and then mined interesting rules for the classification based on membership degrees. Their approach yielded 77.6accuracy. Rutkowski and Cpalka in introduced a new neural-fuzzy structure. called a flexible neural fuzzy inference system (FLEXNFIS). Based on the input and output data, they proposed the parameters of the membership functions and the type of the neuron systems (Mamdani or logical). However, their correct classification percentage on the PID dataset was 78.6%. Davis in developed a fuzzy neural network by using the BK-Square products. This fuzzy neural network was then tested on the PID dataset.

The result of his approach yielded 81.8the results obtained from the Stat Log project when evaluating for many different classification algorithms on the PID dataset showed that their correct classification percentage was less than 78%.The prevalence of type 2 diabetes and intermediate hyper glycemia increases with age. The process from normality to IGT and type 2 diabetes is characterized by progressive insulin resistance or the deterioration of beta cell function (Haffner *et al.,* 1997; Weyer *et al.,*1999). Most of the authors posed their problem as to predict, whether a person would test positive given a number of physiological measurements and medical test results. The prevalence of type 2 diabetes and intermediate hyperglycemia increases with age. The process from normality to IGT and type 2 diabetes is characterized by progressive insulin resistance or the deterioration of beta cell function (Haffner *et al*., 1997; Weyer *et al*., 1999). In DECODE Study Group of papers, 2003a; Qiao *et al*., 2003. non-diabetic Europeans have shown that age is more strongly associated with IGT than with insulin resistance, estimated by homeostasis model assessment. There is substantial literature regarding the efforts made in the field of nonlinear regression. To measure the nonlinearity also to right of entry the adequacy of regression models with their estimation, certain work is available (see e.g., Beale, 1960; Guttman and Meeter, 1965).

Graphical Presentation

Most of the authors calculated the survival curves with median and range majority of these used the Kaplan-Meier method, other methods being life table, actuarial and Nelson estimates (Nelson,1969). graphs on slopes used to connect the points of the survival curve. Censored observations were rarely marked in few papers some authors gave the number of patients at risk at given times and they showed the confidence intervals or standard errors.

DISCUSSION

Most of the deficiencies, described here can be classified as poor reporting rather than errors in methodology, but this should not be taken to suggest that the identified weaknesses as unimportant. Ambiguous descriptions of the methods used makes it difficult or impossible for readers to know what was done. In most of the papers there was an association between response and survival which is generally to be expected, but no information provided about the treatment efficacy. Most of the reviews are concerned with different survival models and their relative application to specific situations that arise in real time scenarios.

This motivated the researcher to study the comparison of Nonparametric, Semi parametric, Parametric and Bayesian Models. The comparison of survival models is an interesting and useful area of research since these applications are mainly focused in medical field and social sciences. Authors should adhere to the advice of the International Committee of Medical Journal Editors (1991) it describes the statistical methods with the details which they can able to a knowledgeable reader with the access on to the original data and to verify the reported results. Poor reporting is one of the failings that is most amenable to improvement. In particular, there was a general tendency to present results. Some of the deficiencies may have been due to limitations of the software used.

REFERENCES

Altman, D.G. 1982. Statistics in medical journals. Stat. Med., 1, 59-71.

Andersen, P.K. 1991. Survival analysis 1982-91: the second decade of the proportional hazards regression model. Stat. Med., 10, 1931-1941.

Armitage, P., Berry, G., Matthews, J.N.S. 2002. Statistical Methods in Medical Research, 4th edn. Oxford, UK: Blackwell Science.

Collett, D. 1994. Modelling Survival Data in Medical Research. Chapman & Hall: London.

Cox, Dr. 1972. Regression models and life-tables (with discussion) J.R. Stat. Soc. B, 39, 86-94.

D.G. Altman, B.L. D.E. Stavola, S.B, Love and K.A. Stepniewska, 1995. Review of survival analyses published in cancer journals. *British Journal of Cancer*, 72, 511-518.

Kalbfleisch J.D., Prentice R.L., 2002. The statistical analysis of failure time data. 2nd ed. Hoboken: John Wiley and Sons.

Kaplan, E.K., MEIER, P. 1958. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.,* 53(282):457-81.

Klein, J.P., Moeschberger, M.L. 2003. Survival analysis: Techniques for censored and truncated data. New York: Springer.

Zelen, M. 1983. Guidelines for published papers on cancer clinical trials: responsibilities of editors and authors. *J. Clin. Oncol.,* 1,164- 169.

APPENDIX

Useful guidelines on presentation of survival analyses

Data Representation

- Describe the recruitment and analysis dates.

- Describe the reason for the sample size.
- Report a summary of follow-up, such as the median and quartiles computed by the reverse Kaplan-Meier method if median was not able to found give reasons such as more number of censorings in the sample size
- Report how many subjects were lost to follow-up and whether, and how, they had been included in the analyses.
- Report the number of events for each end point.

Presentation Methods

- Give a clear definition of each end point being considered i.e. define the time origin, the event of interest and the circumstances where survival times are censored.
- Name the method used for estimating survival probabilities.
- Name any test used in the analyses; justify the test
- The use of weighted log rank tests, with reason why we are using that.
- Report the test for trend when ordered categorical variables are examined.
- When performing univariate or multivariate analyses, report all the covariates examined, their frequency of missing values and the definition of the categories used (if any) whether the covariate is significant or not.
- When Cox regression analyses are performed, describe the criteria used to select the variables in the initial model, the procedure to specify the final model and describe any methods used to assess the model assumptions.
- Name the software used

Presentation of Results

- Give a summary of overall survival: preferably median and/or percent surviving n years.
- If study is a randomised clinical trial, give separate summaries of survival for each treatment group.
- When reporting the results of any test, give the test statistic, the degrees of freedom(when applicable) and the exact P-value.
- When presenting results of a log rank test also report the numbers of observed and expected events in each group (desirable).
- When comparing survival in two or more groups, give an estimate of the survival in each group, e.g. median survival time, survival probabilities for a particular time point, hazard ratio.
- When presenting the results of a Cox regression analysis, report the estimated coefficients (or estimated hazard ratios), measures of their precision (i.e. standard errors or confidence intervals) and/or the associated P-values.
- Do not use crude rates to summarise the data.

Presentation of Graphs

- Use meaningful time intervals.
- Use a step function to join Kaplan-Meier survival estimates.
- Mark the survival time of censored observations (desirable).
- If several curves are reported in the same plot use different lines type (desirable).
- Give number of patients at risk at selected time points (desirable).
- Mark confidence intervals or standard errors for some of the selected time points (desirable).

****