International Journal of Pharmaceutical and Bio-Medical Science

ISSN(print): 2767-827X, ISSN(online): 2767-830X

Volume 01 Issue 06 September 2021

Page No: 72-78

DOI: https://doi.org/10.47191/ijpbms/v1-i6-01 , Impact Factor: 5.374

Survival Analysis: Alternate Approach

Prof. Satyendra Nath Chakrabartty
ORCID ID: 0000-0002-7687-5044
N 304, Vivek Vihar, Sector 82, Noida-201304, Uttar Pradesh, India

ABSTRACT

Existing models of survival analysis dealing with group data have advantages and limitations too. Assumptions of the models need to be verified. Problem arises when one or more assumptions of a model are not satisfied. Methods of survival analysis, inter alia assumes homogeneity of treatment and related factors during the follow-up periods. However, in practice, such assumptions do not hold. The proposed method (Geometric mean approach) is non-parametric, simple and satisfies desired properties from measurement theory angle. Focusing on individual patient, it helps in mathematical diagnosis of disease like cancer of a particular type, disease intensity in terms of the chosen measurable factors/variables, identification of bad prognosis factors of an individual and quantification of progress or deterioration of a patient over time (analogous to hazard function of an individual). The method can help the researchers and practitioners to make meaningful analysis and drawing meaningful conclusions including estimation of hazard function of sample patients without making any assumption. Empirical verifications of the proposed method along with its robustness and estimation of hazard function and clinical validations are proposed as future studies.

KEYWORDS: Assessment of progress; Disease intensity; Geometric mean; Hazard rate; Measurement properties; Prognosis factors.

ARTICLE DETAILS

Published On:
03 September 2021

Available on:
https://ijpbms.com/

1. INTRODUCTION

Survival analysis considers the time between a starting point (e.g. diagnosis of cancer, Bone marrow transplantation (BMT), etc.) and the terminating point (i.e. death or time until an event occurs) and estimate proportion of the persons who survived beyond a specified time interval without occurrence of a particular event. In between the starting point and the terminating event, there are other events (or state) like progress, deterioration, relapse or development of adverse reaction or a new disease entity (like infection). An individual or a sub-group of individuals may transit from one event to the next event or the previous event. However, for all patients, the terminating event may not occur during the period of observations of the study. Thus, number of individuals between two successive events is different.

Two popular approaches to survival analysis are survival function and hazard function. Both depend on time. Large number of factors can influence such functions.

Features of data:

- Time-to-event outcomes defined as the time from the beginning of observation (date of diagnosis/surgery/BMT) to the occurrence of the relevant events (disease recurrence or death) are continuous variables.
- The event of interest i.e. time to death may not occur at the end of follow-up and thus time to event is unknown. This is called censoring. Censoring may also take place in case one or patients are lost during the period of study or they develop a different event which is extremely difficult or impossible for further follow-up.
- A patient can experience relapse and the time of occurrence of the event called “relapse” is not known. Thus, it is difficult to know correctly the time period between a confirmed response and the first relapse of cancer.
- Survival data are longitudinal.
- Usually, survival data are heavily skewed and do not follow Normal distribution. Thus, parametric

statistical analysis assuming normality cannot be applied directly.

- Starting time of all the patients being followed-up are different
- Observation period of the patients who died is different from the same for those who still survive.
- For each time interval, probability (of being a survivor at the end of the interval subject to the condition that the subject was a survivor at the beginning of the interval) is calculated which is a conditional probability.
- There are explanatory variables (like blood pressure) which change in value over time i.e. time-dependent covariates. Inclusion of such time-dependent covariates in a regression analysis needs special considerations since they may be in ratio scale and continuous (like age or tumor size), binary (male or female), unordered categorical (histology) or ordered categorical or ordinal (performance status or stages of FIGO (International Federation of Gynecology and Obstetrics)). However, ordinal responses or FIGO do not satisfy equidistant property i.e. distance between two successive classes (or response categories) is not same and thus, arithmetic aggregation is not meaningful.

2. SURVIVAL FUNCTION AND HAZARD FUNCTION

Considering longitudinal data, two probability functions are estimated viz. survival function and hazard function. Survival function $S(t)$ reflects the probability that a person survives longer than some specified time *t*. Hazard function $\lambda(t)$ gives the probability that an individual who is under observation at a time *t* has an event at that time. Blagoev et al. (2012) described hazard rate (or failure rate) as the rate of occurrence of the event during a given time interval. Thus, $\lambda(t)$ relates to the incidents/event rate, and $S(t)$ reflects the cumulative occurrence or non-occurrence. Note that hazard is a measure of risk. Higher value of hazard in a time-interval implies greater risk (or failure) in that time interval. Results and interpretation depend heavily on the properties of $S(t)$ and $\lambda(t)$ along with associated estimation procedures.

Popular approaches to survival analyses are Kaplan–Meier plots with Log-rank test, and Cox proportional hazards regression.

2.1 Kaplan-Meier estimates and Log-rank test:

Kaplan-Meier method for estimating the survival function is a special case of the life table technique, where the series of time intervals are formed and the death occurs at the beginning of an interval along with the following major assumptions:

- At any time-point, survival prospects of patients who are censored are equal to those who continue to be followed. This assumption is not easy to verify.
- Equal survival probabilities for patients introduced early and late in the study. Large data are required to test the assumption involving data for different subsets.
- Event of interest happens at the time specified. This could be problematic when recurrence/relapse are detected at regular examinations and may result in upward bias of the survival probabilities.

Consider that *n*-patients are being observed at survival times $t_1, t_2, \dots\dots\dots, t_n$ and *r*-deaths. Clearly, $r < n$. Let the ordered times are $t_{(1)} < t_{(2)} <, \dots\dots, < t_{(r)}$.

Let n_j for $j = 1, 2, \dots\dots\dots, r$ denote the number of patients who are alive just before $t_{(j)}$.

Let $d_{(j)}$ be the number of patients who die till $t_{(j)}$.

Then probability of death of a patient in the interval $[\, t_{(j)}, t_{(j+1)})$ can be taken as $\frac{d_{(j)}}{n_j}$ and estimated survival probability in the interval is $\frac{n_j - d_{(j)}}{n_j}$ and Kaplan-Meier estimate of the survival function is $S(t) = \prod_{t_{(j)} \le t} \frac{n_j - d_{(j)}}{n_j}$. For $t < t_{(1)}$, $S(t)$ is taken as 1

Note that $\frac{d_{(j)}}{n_j}$ for different intervals of time reflect hazard function $\lambda(t)$.

Clearly, $S(t) = 1 - \lambda(t)$. Thus, instead of survival function, one can use hazard function.

2,2 Major disadvantages of Kaplan-Meier method:

- Survival probabilities are not always reliable especially for heavy censoring.
- At the end points, the Kaplan-Meier survival curve cannot provide the reliable estimates.
- Kaplan-Meier curve is not a smooth continuous functions and thus, extrapolation is not possible.
- Computation of survival probability at a point can be difficult.
- Censored subjects may reduce the cumulative survival between intervals.
- Reliability of survival curve gets reduced with increase in number of censored patients
- Kaplan-Meier estimators show lower asymptotic efficiency in comparison to same under the parametric setup (Miller, 1981).
- Survival functions based on various factors may differ. Factors which may influence survival functions could be type of disease; type of transplantation (type of BMT from close relatives/siblings or from unknown donor); type of treatments (use of thalidomide, bortezomib, lenalidomide and other classes of medications like: glucocorticoids, DNA alkylating agents, doxorubicin, cisplatinum, etoposide, etc. to patients with Multiple Myeloma(MM) (Rajkumar and Kumar,2016); Raza et.al. 2017), age, gender, socio-economic status of patients and other factors viz.

risk of infection until new cells engraft, development of acute/chronic Graft-Versus-Host Disease (GVHD), etc. Bland and Altman (2004) found that survival curve for anaplastic astrocytoma was higher than the same for glioblastoma. But, that does not mean the population estimate of survival of patients with anaplastic astrocytoma was worse than the patients with glioblastoma.

The KM approach helps to compare two groups at chosen time point(s) but fails to compare total survival experience of the two groups. The Log-rank test compares the entire survival experience between groups and can be taken as a test of identical survival curves.

Null hypothesis of log-rank test, H_0: $Prob(Event\ E)_{Group\ 1} = Prob(Event\ E)_{Group\ 2}$

For each time point, one to compute observed and expected number of deaths in each group assuming no difference between the groups. For example, let number of alive patients at the starting point in Group-1 and Group-2 be $n_1$and n_2 respectively where $n_1 + n_2 = n$. So risk (probability) of death is $\frac{1}{n}$. Expected number of death under the null hypothesis in Group-1 and Group-2 are $n_1.\frac{1}{n} = \frac{n_1}{n}$ and $\frac{n_2}{n}$ respectively. However, with passage of time (weeks/months), n_1, n_2 and hence, n may get reduced depending upon number of deaths in each group. Thus, one can compute expected number of deaths for each group for each subsequent time point whenever death occurs and also total numbers of expected deaths for each group during the entire time period of the study. In case of censoring of survival time, an individual is taken to be at risk of dying in the time point of the censoring but not in subsequent time points. Considering the observed and expected number of deaths for each group at different time points, the χ^2 statistic can be used to test the null hypothesis.

The log-rank test provides a *p*-value for the differences between the groups; it offers no estimate of the actual effect size. In other words, it offers a statistical, but not a clinical, assessment of the factor's impact.

The log-rank test is based on the same assumptions that censoring is not related to prognosis, survival probabilities are the same for subjects introduced early and late in the study, and the events happened at the times specified. Violation of the assumptions can distort the result especially when censoring is more predominant in one group than another or when survival curves intersects. In addition, log-rank test fails to provide estimate of the size of the difference between the groups and thus offers a statistical, but not a clinical, assessment of impact of the factor. Thus, some more assumptions are needed about the data and use of methods like hazard ratio, Cox proportional hazards(PH) model, etc.

2.3 Cox PH model:

The semi-parametric Cox PH model helps to assess effect of factors like treatments, etc. along with the effects of multiple covariates on survival. Survival rate i.e. proportion of subjects surviving at *t*, denoted by $S(t)$ is negatively related to hazard rate i.e. rate of failure or death at a given time (say *t*), denoted by $\lambda(t)$. It can be proved that $\lambda(t) = \frac{f(t)}{S(t)}$ where $f(t)$ is the overall probability density of failing at time *t* and is equal to $f(t) = \frac{\partial S(t)}{\partial t}$ for continuous case. In other words, $\lambda(t)$ and $S(t)$ are related by $\lambda(t) = -\frac{dlog\ S(t)}{dt}$. So, knowledge of either $\lambda(t)$ or $S(t)$ will facilitate determination of the other and either can be used for further statistical analysis.

However, for assessment of covariates as independent/ predictor variables for prediction of survival, selection of covariates needs to be done carefully with chosen strategy avoiding the confounders (Clark et al.2003; Bradburn et al. 2003; Hosmer et al.2008). Combining covariates in interval or ratio scales and others in nominal or ordinal scales may be problematic. Chakrabartty (2020) provided a method of converting discrete ordinal scores to continuous scores in a desired score range following normal distribution.

Assumptions of Cox PH regression include:

(i) A common baseline hazard function for all patients

(ii) Effect of a covariate is same at all the time points

(iii) Log-hazard is linearly related withthe covariates.

Under these assumptions, Cox PH regression model involving *k*-covariates $X_1, X_2, \ldots\ldots, X_k$ is given by $\lambda(t) = \lambda_0 e^{\sum_{i=1}^{k} \beta_i X_i}$ where $\lambda_0$the base line hazard (estimated in non-parametric fashion) and β_i is the coefficient of X_i reflecting effect size of the *i*-th covariate.

$\beta_i > 0 \Rightarrow X_i$ is positively related to hazard i.e. increase in X_i will increase $\lambda(t)$ and decrease length of survival. e^{β_i} is the hazard ratio (HR).

Thus, the covariates can be ranked in terms of values of beta-coefficients or equivalently by HR. In cancer studies, $e^{\beta_i} > 1 (\text{or } \beta_i > 0)$ means X_i is a bad prognostic factor. It is desirable to have more $\beta_i's < 0$.

Clearly, ratio of two hazard functions of two groups is $\frac{\lambda_1(t) for\ Group\ 1}{\lambda_2(t) for\ Group\ 2}$ = Constant.

Note that the Cox regression model does not involve the residual error term i.e. absolute difference between actual and predicted value of hazards. This is unlike the regression equation where residual error ϵ_i is assumed to follow $N(0, \sigma^2)$. This is due to the fact that Cox regression model introduces time-dependent censored variables for cases when no analyzed end point has occurred during the follow-up.

Logistic regression can help to investigate relationship between various disease events and the risk factors. But the Cox model is preferred over the logistic model, since the later ignores survival time and censoring information (Fabsic et al.2011)

Violation of the assumption of proportionality of the hazards of the Cox regression model, may give distorted results. To investigate effect of gender on survival time of 48 patients who were diagnosed and died of multiple myeloma (MM), Mamudu and Tsokos (2020) used 16 identified risk factors (11 continuous risk factors, and 5 categorical risk factors). *p-value* of the log-rank test was 0.45 implying rejection of the null hypothesis of no difference with respect to gender. Similarly, in multivariable Cox regression analysis, Babińska et al. (2015) found violation of the proportional hazard assumption for covariates like concentrations of homocysteine and sodium which increased risk of death and recommended need to verify this assumption of proportionality. In case two hazard functions corresponding to two groups cross, it indicates violation of the assumption of proportionality of Cox PH model. If one draws regression line of residuals across the time-axis and if the obtained line is horizontal, constant proportionality of hazard is ensured (Abeysekera and Sooriyarachchi, 2009). Global goodness-of-fit test proposed by Schoenfeld is more rigors for testing the PH assumption (Schoenfeld, 2017).

In case of non-satisfaction of the assumption of proportionality, the Cox regression model needs to be modified to Cox stratified regression model (allows inclusion of stratification (categorization) of a variable not satisfying the assumption of the proportionality and does not require to define the interaction relationship between a stratified variable and the observation time) (Ata and S¨ozer, 2007).

2.4 Parametric models:

Parametric PH models like Exponential, Weibull, Gompertz models select different hazard functions which are assumed to follow specific probability distribution and thus, differ from Cox semi-parametric PH model. Parametric models usually have smaller standard error and thus, are considered as more efficient. However, verification of the assumption of hazard function following a specific probability distribution may be difficult.

Accelerated failure time (AFT) model with focus on survival function differs from Cox models. Parametric AFT model uses regression of logarithm of the survival time over the covariates. It can be an alternative to the Cox model (Wei, 1992). Commonly used distribution in AFT model is log-logistic distribution. However, other distributions like log-normal, gamma and inverse Gaussian distributions, can also be suitable for AFT models. AFT allows derivation of a time ratio, which is easier to interpret than a ratio of two hazards. However, assumed distributions of parametric Cox models, AFT etc. are required to be verified with appropriate statistical tests. Uses of AFT models in medical research are less popular (Kay and Kinnersley, 2002)

2.5 Multivariate Linear Regression:

For predicting survival time (*t*) with known values of 16 chosen risk factors and their pairwise interactions, Mamudu and Tsokos(2020) fitted multivariate linear regression equation of the form $t_i = \alpha + \sum_{i=1}^{k} \beta_i X_i + \sum_{i \neq j=1}^{k} \rho_{ij} X_i X_j + \epsilon_i$ where ρ_{ij} is the coefficient parameter of interaction between the *i-th* and *j-th* risk factors, and ϵ_i is the error to predict survival time for the *i*-th patient, *i*=1, 2, …….48, and *k*=16, the number of risk factors.

However, application of such model requires verification of the following assumptions of multivariate linear regression:

- *Linearity*: usually by correlation between the response and the continuous risk factors. However, high value of correlation between X and Y may not imply linearity. Chakrabartty (2020 b) gave examples of $r_{XY} \geq 0.9$ where X takes values 1, 2, …., 30 and $Y = X^2$ or X^3 or log_{10}^X and suggested for testing of significance of standard error of prediction than testing significance of correlation coefficient.
- *Normality and Homoscedasticity*: Residual error to follow $N(0, \sigma^2)$
- *Multicollinearity*: If a pair of risk factors is highly correlated, it may amount to double counting. Presence of multicollinearity, if any needs to be avoided. High value of the variance inflation factor (VIF) indicates existence of multicollinearity. Researchers differ in their approaches, if VIF is high.
- *No autocorrelation*: Autocorrelation is the degree of similarity between the values of the same variables over successive time intervals. Presence of autocorrelation in the residuals of a model implies that the model is not sound. Checking whether residual errors are independent and uncorrelated, usually Durbin-Watson test is used with null hypothesis of no autocorrelation
- *Goodness of fit*: High value of R^2 defined as $1 - \frac{SS_{Residual}}{SS_{Total}}$ implies goodness-of-fit of a statistical model. It increases with increase in number of predictor variables. Better measure is $R^2_{Adjusted} = 1 - \frac{SS_{Residual}/df_{Residual}}{SS_{Total}/df_{Total}}$

2.6 Cumulative Incidence in Competing Risks:

Competing risks are common in medical research. Examples of competing risks are: Cancer-related mortality and deaths due to reasons other than cancer; death due to cardiovascular causes and death attributable to non-cardiovascular causes; etc. Time-to-event analyses without considering competing risks can lead to biased estimates of risk for patients with multi-morbidity and belonging to higher age category (Abdel-Quadir et al. 2018). Existing methods of survival analysis like Kaplan-Meier estimation of cumulative incidence, log-rank test for comparing cumulative incidence curves, Cox model for the assessment of covariates, etc. may lead to biased results in analysis of competing risks data (Kim, 2007). The author found that analysis of "death" ignoring CR resulted in erroneous results in terms of high or low magnitude of error depending on high or low CR

incidence respectively and suggested analysis of joint events as a single end point along with CR, which offers a comprehensive approach to addresses general and specific research questions like difference between cumulative incidence of TRM (CIT) and cumulative incidence of relapse (CIR) in myeloablative versus non-myeloablative HSCT studies.

CR regression analysis helps in identification of risk factors associated with each competing risk and also to find reasons of difference in the cumulative incidence curves. CR regression analysis can be carried out by models given by Fine and Gray (1999) (based on proportional hazards) and Klein and Andersen (2005) (based on pseudo values emerging from a jackknife statistic). Empirically, the above said two methods were in close agreement. However, the issues relating to sample size and power calculation in CR data are needed to be addressed.

3. PROSED METHOD

Assuming there are *n*- key non-nominal variables, observed values of the *i*-th person can be presented as a vector $\boldsymbol{X_i} = (X_{i1}, X_{i2}, \dots, X_{in})^T$ and the corresponding value of standards (lower value) as another vector $\boldsymbol{X_0} = (X_{01}, X_{02}, \dots, X_{on})^T$ where each variable has been transformed to be positively related to cancer. For example, variables like Platelet count, WBC count, % Myeloid cells in peripheral blood, etc. whose lower value indicates higher risk to cancer, reciprocal of such variables are taken. For example, standard for Platelet count may be taken as $\frac{1}{450,000}$ to $\frac{1}{140,000}$ cells/mcL^{-1} or $\frac{1}{450}$ to $\frac{1}{140}$ thousand/mm^3. For variable like Basophils (DLC), where instead of a range, a single value is given in the reference range; an agreed particular value may be taken as the standard.

Consider the unit free ratios $Y_{i1} = \frac{X_{i1}}{X_{01}}, Y_{i2} = \frac{X_{i2}}{X_{02}}, \dots, Y_{in} = \frac{X_{in}}{X_{on}}$ for the *i*-th person. Each ratio is positive. Value of a ratio may be $<$ or ≥ 1. Since, all the variables have been converted to be positively related to cancer, $Y_{ij} > 1 \Longrightarrow$ the *i*-th person exceeded the standard value of the *j*-th variable and run the risk of cancer w.r.t. the *j*-th variable. If $Y_{ij} > 1$ for more than one value of *j*, it would imply the person's risk of cancer with respect to more than one variable. If Y_{ij} is close to1 from left for the $j-$th variable, the *i*-th person may not be diagnosed as cancer patient but may be taken as a potential cancer patient with respect to the *j*-th variable.

However, quantification of *intensity of cancer* of the *i*-th person could be assessed by finding similarities or deviations between the values of the vector $\boldsymbol{X_i}$ and $\boldsymbol{X_0}$ by the geometric mean of the unit free values of $Y_{i1}, Y_{i2}, Y_{i3}, \dots, Y_{in}$ for the *i*-th person.

Thus, cancer intensity (CI_i) of the *i*-th person is

$$CI_i = \sqrt[n]{\prod_{j=1}^{n} Y_{ij}} \qquad (1)$$

Alternatively, avoiding the *n*-th root, CI_i can be taken as

$$CI_i = \prod_{j=1}^{n} Y_{ij} \qquad (2)$$

Equation (2) is preferred for little lesser computations.

If value of $CI_i > 1$, the person can be diagnosed as cancer patient and the $Y'_{ij}s$ exceeding one are the bad prognostic factors for the *i*-th person requiring attention to decide individual patient care and choice of treatment. Bad prognosis factors for a patient may differ depending on types of disease and his/her disease intensity.

The proposed index CI_i reflects disease-status considering all the chosen factors or cancer intensity of the *i*-th person and thus, facilitates ranking a group of patients in terms of cancer intensity. Contribution of the *j*-th factor to disease intensity of the *i*-th patient is given by $\frac{Y_{ij}}{CI_i}$. CI_i can be multiplied by 100 to give percentage figures. The simple, unit free, continuous measure with monotonic property satisfies all the scientific aggregation rules, given by Ebert and Welsch, (2004) and has the following properties:

i) Consider all chosen variables and depicts disease/cancer intensity of the *i*-th person with respect to the standards.

ii) Not affected by change of scale

iii) Assess relative importance of the factors to disease intensity of a patient.

iv) Can be applied also for skewed data, even when the chosen variables are correlated amongst themselves or with the disease intensity (confounding effect).

v) Presence of extreme values (outliers) cannot affect the measure much and thus no bias for measuring disease intensity of a patient.

vi) Low value of one key variable does not get compensated by high values in another key variable.

vii) Facilitates formation of chain indices, that is $CI_{20} = CI_{21}.CI_{10}$ where "0" denotes the first time period (or base period), *i*-th patient was diagnosed; 1 and 2 denote the subsequent time- periods when *CI* were measured for the patient.

viii) Possible to obtain graph of CI_i over a period to reflect path of improvement and relapses experienced by the patient over time.

3.1 Advantages:

3.1.1 Estimation and testing: Sample GM can be computed using $logGM = \frac{1}{n}\sum logY_{ij}$. Population estimate of GM can be taken as the sample GM for large or moderately large data involving sample size of *N*. Alf and Grossberg (1979) have given estimate of standard error of the GM and confidence interval of GM. Testing of null hypotheses $H_0: GM_1 = GM_2$ can be performed by conventional $t-$tests on the logarithms of the observations.

3.1.2 Classification: CI_i's can be used for undertaking Cluster analysis and classification of cancer patients with respect to cancer intensity, separately for each cancer type.

However, class membership of a patient may get changed subsequently with changed value of CI_i due to cares and

Corresponding Author: Prof. Satyendra Nath Chakrabartty

treatments. Such analysis with large data emerging from representative samples under each cancer type may help to find norms (class boundaries) for the classes.

3.1.3 Assessment of progress: Let CI_{i_1} be the cancer intensity of the *i*-th person at time-point 1, i.e. the first time when the patient was examined and treatment started. So,

$$CI_{i_1} = \prod_{j=1}^{n} Y_{ij_1} = \prod_{j=1}^{n} \frac{X_{ij_1}}{X_{oj}} \quad (3)$$

CI_{i_1} is the baseline status of disease-characteristics of the *i*-th patient. Subsequently, cancer intensity of the same patient are assessed at different time periods CI_{i_t} for *t*= 1, 2, 3, … , and so on. With passage of time, CI_{i_t} may show zigzag pattern as many patients with Multiple Myeloma (MM) respond to initial therapy but recorded relapse (Sonneveld et al. 2017).

$CI_{i_t} < CI_{i_{(t-1)}}$ indicates improvement of the patient during the *t*-th time from the (*t*-1)-th time. Similarly, $CI_{i_t} > CI_{i_{(t-1)}}$, imply deterioration of patient during the same period. Thus, negative value of $[CI_{i_t} - CI_{i_{(t-1)}}]$ will reflect reduction in disease intensity or progress of the *i*-th patient at *t*-th time from the (*t*-1)-th time. The same may also be used to find short-term effect of surgery or BMT or a clinical trial. The critical variables where deterioration took place can be observed by comparing the values of Y_{ij} for the period *t* and (*t*-1), that is those component variables for which Y_{ij_t} exceeds $Y_{ij_{(t-1)}}$. Extent of deterioration in the identified variables can be assessed by difference of values of corresponding $X'_{ij}s$.

Alternatively, progress of a patient during *t*-th time point over (*t*-1)-th period can be reflected by the ratio $\frac{CI_{i_t}}{CI_{i_{(t-1)}}}$. Value of the ratio exceeding unity will indicate improvement in the *t*-th period from (*t*-1)-th time point. Note that the ratio

$$\frac{CI_{i_t}}{CI_{i_{(t-1)}}} = \frac{\prod_{j=1}^{n} Y_{ij_t}}{\prod_{j=1}^{n} Y_{ij_{(t-1)}}} = \frac{\prod_{j=1}^{n} \frac{X_{ij_t}}{X_{oj}}}{\prod_{j=1}^{n} \frac{X_{ij_{(t-1)}}}{X_{oj}}} \quad (4)$$

The equation (4) helps to find progress of a patient from the beginning (i.e. time zero) since the measure CI_{i_t} facilitates formation of chain indices.

A decreasing graph of CI_{i_t} and *t* implies that the *i*-th patient is improving over time and an increasing graph of CI_{i_t} and *t* will indicate the reverse. Attempt can be made to find a small interval of values of CI_{i_t} for each cancer type which may be associated with *Stage IV cancer* or *metastatic cancer*.

3.1.4 Hazard function: The graph of CI_t emerging from (4) is more akin to hazard function $\lambda(t)$ since $\frac{X_{ij_t}}{X_{oj}}$ quantifies extent of *j*-th hazard for the *i*-th patient at *t*-th time. For a sample of size *N*, there are *N* points using $logGM_i = \frac{1}{n}\sum_{j=1}^{n} log \frac{X_{ij}}{X_{oj}}$ *i*=1, 2, ……, *N*.

Further observations of patients in subsequent periods, will generate additional points. The frame allows fitting of multiple linear regression of the form $\log GM = \alpha + \sum \beta_i log X_i + \epsilon$ which is an alternative approach to estimate hazard function of patients without any assumption. For completeness of follow-up data, one may ignore the patients who discontinued or died early. However, this needs to be compared with Cox hazard function empirically.

4. DISCUSSION

Non-parametric Kaplan-Meier survival curve is popular but has several limitations and cannot assess the relationship of survival with the explanatory variables.

Existing semi-parametric and parametric models dealing with group data have advantages and limitations too. Assumptions of the models need to be verified. Major question to address is 'what to do if one or more assumptions are not satisfied'.

Methods of survival analysis, inter alia assumes homogeneity of treatment and related factors during the follow-up period (Clark et al. 2003). However, in practice, such assumptions do not hold.

The proposed method (Geometric mean approach) is non-parametric, simple and satisfies desired properties from measurement theory angle. Strictly speaking, this is not model driven. Focusing on individual patient, it helps in mathematical diagnosis of disease like cancer of a particular type, disease intensity in terms of function of geometric mean of the chosen measurable factors/variables, identification of bad prognosis factors of an individual and quantification of progress or deterioration of a patient over time (analogous to hazard function of an individual). The method can help the researchers and practitioners to make meaningful analysis and drawing meaningful conclusions including estimation of hazard function of sample patients without making any assumption.

However, empirical verifications of the proposed method, its robustness and estimation of hazard function and clinical validations are proposed as future studies.

REFERENCES

I. Abdel-Qadir,H.; Fang,J.; Lee, DS.; Tu, Jack V.; et al. (2018): Importance of Considering Competing Risks in Time-to-Event Analyses, *Circulation: Cardiovascular Quality and Outcomes*, 11 (7); 1 - 11. https://doi.org/10.1161/CIRCOUTCOMES.118.004580

II. Abeysekera, W. W. M., & Sooriyarachchi, M. R. (2009): Use of Schoenfeld's global test to test the proportional hazards assumption in the Cox proportional hazards model: an application to a clinical study. *Journal of the Naional Science Foundation of Sri Lanka*, 37(1), 41–51.

III. Alf, Edward F. and Grossberg, John M. (1979): The geometric mean: Confidence limits and significance tests.*Perception& Psychophysics*, 26 (5); 419–421

IV. Ata, N., & S¨ozer, M. T. (2007): Cox Regression Models with nonproportional Hazards

V. applied to Lung Cancer Survival Data. *Hacettepe Journal of Mathematics Statistics*, 36(2), 157–167.

VI. Babińska1, M.; Chudek,J.; Chełmecka, E.; Janik, M.; Klimek, K. and Owczarek, A. (2015): Limitations of Cox Proportional Hazards Analysis in Mortality Prediction of Patients with Acute Coronary Syndrome, *Studies in Logic, Grammar and Rhetoric*. 43. 10.1515/slgr-2015-0040. DOI:10.1515/slgr-2015-0040

VII. Blagoev KB, Wilkerson J, Fojo T. (2012): Hazard ratios in cancer clinical trials–a primer. *Nat Rev Clin Oncol*. 2012;9:178–183

VIII. Bland JM, Altman DG.(2004): The logrank test. *BMJ*. ; 328(7447): 1073 10.1136/bmj.328.7447.1073

IX. Bradburn MJ, Clark TG, Love SB, Altman DG. (2003): Survival analysis part III: multivariate data analysis—choosing a model and assessing its adequacy and fit. *Br J Cancer*. 2003;89:605–611

X. Chakrabartty, SN. (2020): Improve Quality of Pain Measurement, *Health Sciences*, Vol.1, ID 259, 1 – 6. DOI: 10.15342/hs.2020.259

XI. Chakrabartty, Satyendra Nath (2020b): Better Use of Scales as Measuring Instruments in Mental Disorders. *Journal of Neurology Research Review & Reports*. SRC/JNRRR-128. DOI: https://doi.org/10.47363/ JNRRR/2020-(2)128

XII. Clark TG, Bradburn MJ, Love SB, Altman DG. (2003): Survival analysis part I: basic concepts and first analyses. *Br J Cancer*. 2003; 89:232–238.

XIII. Ebert, U. and H. Welsch, (2004): Meaningful Environmental Indices: a Social Choice Approach. *Journal of Environmental Economics and Management*, 47: 270-283.

XIV. Fabsic P, Evgeny V, Zemmer K, editors. Seminar in Statistics: Survival Analysis Presentation 3: The Cox proportional hazard model and its characteristics. Zurich: 2011.

XV. Fine JP and Gray RJ. (1999): A proportional hazards model for the sub distribution of a competing risk. *J Am Stat Assoc*1999; 94: 496 - 509.

XVI. Hosmer DW, Lemeshow S, May S. (2008): Model development. *In: Applied Survival Analysis*:2nd Ed Hoboken, NJ: John Wiley & Sons; 132–168.

XVII. Kay R, and Kinnersley N. (2002): On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: a case study in influenza. *Drug Inf. jr.*;36: 571–579.

XVIII. Kim, Haesook T. (2007): Cumulative Incidence in Competing Risks Data and Competing Risks Regression Analysis, *Clin Cancer Res*;559 13(2), 559 – 565. doi: 10.1158/1078-0432.CCR-06-1210.

XIX. Klein JP, Andersen PK. (2005): Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*;61: 223 - 229.

XX. Mamudu, L and Tsokos, CP. (2020): Data-Driven Statistical Modeling and Analysis of the Survival Times of Multiple Myeloma Cancer, *Health Science Journal*, Vol. 14 No. 1: 693; DOI: 10.36648/1791-809X.14.1.693

XXI. Miller, R.G. (1981): *Survival Analysis*, John Wiley and Sons, New York.

XXII. Rajkumar SV, Kumar S (2016): Multiple myeloma: diagnosis and treatment. *Mayo Clin Proc* 91: 101-119.

XXIII. Raza S, Safyan RA, Rosenbaum E, Bowman AS, Lentzsch S (2017): Optimizing current and emerging therapies in multiple myeloma: a guide for the hematologist. *Ther Adv Hematol* 8: 55-70.

XXIV. Sonneveld P, De Wit E, Moreau P (2017): How have evolutions in strategies for the treatment of relapsed/refractory multiple myeloma translated into improved outcomes for patients? *Crit Rev Oncol Hematol* 112: 153-170.

XXV. Wei, L. J. (1992): The accelerated failure time model: A useful alternative to the cox regression model in survival analysis, *Statistics in Medicine*, Vol.11, Issue 14-15, 1871-1879. https://doi.org/10.1002/sim.4780111409