Journal of Probability and Statistical Science
22(1), 1-15 Sep. 2024

On the Estimation of Survival Rate Using Ranked Set Sampling Design

Etebong P. Clement
Department of Statistics
University of Uyo, Nigeria

Rose C. Akpan
Local Government Pensions Board
Uyo, Akwa Ibom State, Nigeria

ABSTRACT

The Nigerian pensioners have high expectations from the government to ensure an effective implementation of pension regulations existing in the country. These expectations arise from the need to have sustainable standard of living in retirement and their benefits paid as at when due. However, the government does not know how many retirees will be alive at a particular period of time to draw fund from the pension account. This calls for regular verification of pensioners by Pension Transitional Arrangements Directorate (PTAD) at national level, office of the accountant general for State level and Local Government Pensions Board for Unified Local Government Council. This process exposed pensioners to stress related problems and some pensioners even die during such verification exercise. Again, it gives rise to varying set of problems that limit the capacity of the stakeholders within Nigeria pension industry to meet pensioner's expectations. Survival rate of these retirees is key important statistic to the government for adequate preparation for their terminal benefits and to provide other social packages for this vulnerable group of people. However, this statistic is lacking in the State. Consequently, this work intends to address this problem by using the theory of Ranked Set Sampling (RSS) for Survival analysis to estimate the Survival rate of the retirees which will help the government in making adequate budgetary provisions to the Nigeria Pension Industry. Analysis and evaluation are presented.

Keywords: Ranked Set Sampling, Survival Analysis, Kaplan-Meier estimator, Cox proportional Hazard model and Hazard ratio.

1. Introduction

The modeling of time to event data is an important topic with many applications in diverse areas. The collective of methods to analyze such data are called survival analysis, event history analysis or duration analysis (Emmert–Streib and Dehmer, 2019). Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death which is very common in medical field, death after retirement as in the work of Bamia *et al*. (2008), relapse and recurrence (Fields *et al*. (2011), agitation attack (Lesem *et al*.,(2011), graduation (Min *et al*. (2011), malfunctioning of device (Richardeau and Pham, 2012) and bankruptcy (Daepp *et al*. (2015) or any designated experience of interest that may happen to an individual.

The concept of RSS was developed by McIntyre (1952) to estimate mean pasture yields. Takahas1 and Wakimoto (1968) extended the theory of RRS under presumption of perfect ranking. Other notable scholars in RSS include: Gaur *et al*. (2013) considered an RSS approach to the multiple sample scale problem. Chen *et al*. (2013) extended RSS to Moving Extreme Ranked Set Sampling (MERSS) for estimation of scale parameter for scale distribution, Singh *et al*. (2014) proposed an estimator for population mean and ranking of the elements observed on the basis of auxiliary variable. Al-Omari and Bouza (2014) used RSS design to assess the impact of developmental programme and discussed the procedure of RSS design. Khan and Shabbir (2015) suggested a class of Hartley-Ross type unbiased estimator in RSS.

- Received October 2023, in final form August 2024.
- Clement, Etebong P. (corresponding author) and Akpan, Rose C. are affiliated with the Department of Statistics, University of Uyo, Nigeria.
epclement@yahoo.com

In RSS the population is divided into a simple random sample of size k, each unit is rated according to subjective criteria. The smallest unit in the sample is measured and the remaining units are eliminated. After ranking each unit according to the same criteria, a second simple random sample of size k is chosen from the population and the second smallest unit is then measured and the remaining units are discarded. This process is repeated until the ordered units are measured.

As seen in the work of Zhang (2016), instead of focusing on the time (how long) a subject can survive, survival analysis examines the probability of an event given subjects who are under observation at that survival time. With this, survival times and survival probability can be estimated without bias given that subjects under observation are true representatives. For this paper event of interest is survival after retirement. Survival rate for these Retirees is key important statistic to the government for adequate preparation for their terminal benefits and to provide other social packages for this vulnerable group of people. On the whole, this statistic is lacking in the State. Consequently, this work intends to address this problem by introducing the theory of Ranked Set Sampling (RSS) for Survival analysis to estimate the Survival rate of the retirees which will help the government in making adequate budgeting provisions to the Nigeria Pension Industry.

2. Materials and methods

2.1 Basic Notations and Definitions

Let $X_1, X_2, X_3, \dots, X_n$ be a random sample of size n from a random variable X with probability density function $\mathbf{f(x)}$ and finite mean (μ) and variance (σ^2). Let $X_{i(1)}$ be the first order statistic from the set $(X_{i1}, X_{i2}, \dots, X_{in})$ which represents the ith random sample of size n. For convenience $X_{i(1)}$ is also written as $X_{(i,n)}$to denote the *ith* order statistic from the *ith* set of n observations with mean $\mu_{(i,n)}$. Let $X_{(i,n)j}$ denote the ith order statistic from the *ith* sample of size n in the *jth* cycle ($\mathrm{j} = 1, 2, \dots, \mathrm{m}$).

According to Wolfe (2012) set size plays a critical role in the performance of any RSS procedure. For a given set size k, each measured ranked set sample observation utilizes additional information obtained from its ranking relative to $k-1$ other units from the population. With perfect rankings this additional information is clearly an increasing function of k Thus, with perfect rankings, is it good to take set size k to be as small as possible, the larger k is, the more likely to experience ranking errors.

Survival time data are analyzed with the use of special techniques and the underlying assumptions taken into account. According to Zwiener *et al*. (2011) survival times are analyzed with the Kaplan-Meier method which yields two measures of interest: survival rates and the median survival time. The log-rank test is used to compare survival times across treatment groups while Cox regression model is used to test the effect of other independent variables on the survival time.

2.2 Estimation of Population of interest

The natural ranked set sample estimator, $\hat{\mu}_{RSS}$, for the population mean μ based on the ranked set sample $\left(X_{(1)}, \dots, X_{(k)1}; X_{(1)2}, \dots, X_{(k)2}; \dots; X_{(1)M}, \dots, X_{(K)M}\right)$ is simply the average of the sample observations.

The unbiased estimator of the population mean is determined using (Wolfe, 2012) as given in Equation (1)

$$\hat{\mu}_{RSS} = \bar{X}_{RSS} = \sum_{j=1}^{m}\sum_{i=1}^{k}\frac{X_{[i]j}}{km} \quad (1)$$

The balanced RSS estimator $\hat{\mu}_{RSS}$ in equation (1) is also an unbiased estimator for the population mean μ regardless of whether the judgment rankings are perfect or imperfect. For simplicity, let consider only the case of a single cycle ($m = 1$), so that the total sample size n is equal to the set size k. Under the assumption of perfect rankings, the RSS observations can be represented for this setting by $X^*_{(1)}, \ldots, X^*_{(k)}$ where the k variables are mutually independent and $X^*_{(i)}$, $i = 1,2, \ldots, k$ is distributed like the ith order statistic for a random sample of size k from a continuous distribution with distribution function F and density f. Thus (1) becomes

$$\hat{\mu}_{RSS} = \bar{X}_{RSS} = \frac{1}{\mathrm{K}}\sum_{i=1}^{k} X^*_{(i)} \quad (2)$$

Taking expectation of (2) gives

$$E[\hat{\mu}_{RSS}] = E[\bar{X}_{RSS}] = \frac{1}{k}\sum_{i=1}^{k} E\left[X^*_{(i)}\right] \quad . \quad (3)$$

$$E\left[X^*_{(i)}\right] = \int_{-\infty}^{\infty} x \frac{k}{(i-1)!(k-i)!} [F(x)]^{i-1} [1-F(x)]^{k-1} f(x)dx \quad (4)$$

Substituting (4) in (3) gives

$$E[\bar{X}_{RSS}] = \frac{1}{k}\sum_{i=1}^{k}\{\int_{-\infty}^{\infty} kx\binom{k-1}{i-1} [F(x)]^{i-1} [1-F(x)]^{k-1} f(x)dx\} \qquad i = 1,2, \ldots, k$$

$$E[\bar{X}_{RSS}] = \int_{-\infty}^{\infty} xf(x)\{\sum_{i=1}^{k}\binom{k-1}{i-1} [F(x)]^{i-1} [1-F(x)]^{k-i}\}dx$$

$$E[\bar{X}_{RSS}] = \{\int_{-\infty}^{\infty} xf(x)[\sum_{i=1}^{k}\binom{k-1}{q} [F(x)]^{q} [1-F(x)]^{k-i}]\}dx$$

Since $\sum_{q=0}^{k-1}\binom{k-1}{q} [F(x)]^{q} [1-F(x)]^{(k-i)-q} = 1$

and $q = i - 1$

$$E[\mu_{RSS}] = E[\bar{X}_{RSS}] = \int_{-\infty}^{\infty} xf(x)dx = \mu. \quad (5)$$

Following from (2)

$$Var[\bar{X}_{RSS}] = \frac{1}{k^2}\sum_{i=1}^{k} Var(X^*_{(i)}) \quad (6)$$

Let $\mu^*_{(i)} = E[X^*_{(i)}]$, for $i = 1, \ldots, k$, then $E\left[(X^*_{(i)} - \mu)^2\right] = E\left[\left(X^*_i - \mu^*_{(i)} + \mu^*_{(i)} - \mu\right)\right]^2$

$$E\left[[X^*_{(i)} - \mu]^2\right] = Var\left(X^*_{(i)}\right) + \left(\mu^*_{(i)} - \mu\right)^2$$

$$Var\left(X^*_{(i)}\right) = E\left[[X^*_{(i)} - \mu]^2\right] - \left(\mu^*_{(i)} - \mu\right)^2 \quad (7)$$

substituting (7) in (6) gives (8)

$$Var[\bar{X}_{RSS}] = \frac{1}{k^2}\sum_{i=1}^{k} E[(X^*_{(i)} - \mu)^2)] - \frac{1}{k^2}\sum_{i=1}^{k}(\mu^*_{(i)} - \mu)^2 \; . \quad (8)$$

Similarly,

$$\sum_{i=1}^{k} E(\mathrm{X}^*_{\mathrm{i}} - \mu)^2) = \sum_{i=1}^{k}\int_{-\infty}^{\infty} k(x-\mu)^2 \binom{k-1}{i-1} [F(x)]^{i-1} [1-F(x)]^{k-\mathrm{i}} f(x)dx$$

$$\sum_{i=1}^{k} E(\mathrm{X}^*_{\mathrm{i}} - \mu)^2) = k\int_{-\infty}^{\infty}(x-\mu)^2 f(x)\{\sum_{i=1}^{k}\binom{k-1}{i-1} [F(x)]^{i-1}\left[1-F(x)\right]^{k-i}\}dx \quad (9)$$

$$\sum_{i=1}^{k} E[(X^*_{(i)} - \mu)^2)] = k\int_{-\infty}^{\infty}(x-\mu)^2 f(x) = \mathrm{k}\delta^2. \quad (10)$$

Substituting (10) in (8), it follows that

$$Var[\bar{X}_{RSS}] = \frac{1}{k^2}\{\mathrm{k}\delta^2 - \sum_{i=1}^{k}(\mu^*_{(i)} - \mu)^2)\} = \frac{\delta^2}{k} - \frac{1}{k^2} \quad (11)$$

Thus, both $\hat{\mu}_{SRS}$ *and* $\hat{\mu}_{RSS}$ are unbiased estimator for population mean.

Working with a finite population, the formula for determining the sample size is given by Saxen (2018) as:

$$\mathrm{n} = \frac{\mathrm{N}}{(1+\mathrm{Ne}^2)} \quad (12)$$

where n is the corrected sample size, N is the population size, and e is the Margin of error (MoE) given by $\mathrm{e} = 0.05$ based on the research condition by Saxen (2018).

For this study$\mathrm{N} = 1{,}647$; hence

$$n = \frac{1647}{(1+1649(0.05)^2)} \approx 322 \quad (13)$$

2.3 Procedures for selection of sample size based on RSS design

RSS of size n = 320, for the purpose of balanced RSS design, the set size k = 5, m = cycles and N = 1,647 The process proceeded by numbering the retirees names from 1 to 1,647 in the list, with selection of five numbers randomly from four digits random numbers table and put it in a set and collect another SRS of five numbers independent of the first set and placed in another set till five sets independent of each selection without replacement is completed. The smallest number in the first set is considered as the 1st item in RSS. The second smallest in the second set is considered as the 2nd item in RSS. The third smallest in the third set is considered as the 3rd item in RSS. The fourth smallest in the fourth set is considered as the 4th item in RSS and the largest unit is ranked as the 5th item in RSS from the last set and this complete the first cycle. The total of 25 SRS set of numbers are needed for complete one cycle in this very design. The process is repeated 64 times for complete 64 cycles as shown in the matrix below.

$$\begin{bmatrix} X_{1[1]}X_{1[2]} & \cdots & X_{1[5]} \\ X_{2[1]}X_{2[2]} & \cdots & X_{2[5]} \\ X_{3[1]}X_{3[2]} & \cdots & X_{3[5]} \\ \vdots & \vdots \quad \cdots & \vdots \\ X_{64[1]}X_{64[2]} & \cdots & X_{64[5]} \end{bmatrix}$$

Each row represents the set size and the cycle, each cycle produced ith judgment order statistic denoted as $X_{[i]};\ i = 1, \ldots, 5$,. This scheme, RSS produced the required Sample size.

2.4 Kaplan –Meier (KM) Estimator

The Kaplan–Meier (KM) estimator of a survival function S_{KM} (t) according to Emmert-Sterib and Dehmer (2019) is defined as:

$$S_{km}(t) = \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right) \quad (14)$$

Survival data contains two components; survival time and event status. In this study KM estimator is developed using RSS design. KM estimator is a non-parametric method. It measures the probability that a person survives longer than a specific time, which is fundamental in survival analysis. Under RSS design, KM in Equation (14) can be rewritten as in Equation (15)

$$\hat{S}_{RSS}(t) = \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right) \quad (15)$$

The Kaplan–Meier (KM) estimator of a survival function S_{KM} (t) according to Emmert-Sterib and Dehmer (2019) is defined as:

$$S_{km}(t) = \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right) \quad (14)$$

Survival data contains two components; survival time and event status. In this study KM estimator is developed using RSS design. KM estimator is a non-parametric method. It measures the probability that a person survives longer than a specific time, which is fundamental in survival analysis. Under RSS design, KM in Equation (14) can be rewritten as in Equation (15)

$$\hat{S}_{RSS}(t) = \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right) \quad (15)$$

This estimator holds for all t $>$ 0 and it depends only on two variables, $n_i$and d_i Where $\hat{S}_{RSS}$ indicates that this estimator is under RSS design, n_i: number at risk at time t_i, d_i: number of subjects that do not experience the events of interest at time t_i. Here, n_i corresponds to the number of subjects present at time t_i generated through RSS design. In contract, subjects that are censoring are no longer present. This estimator is considering only events i that occur before time t, that is,$t_i < 1$. Hence, the survival curve S_{RSS} (t) for time t under this design considers all events that happened before t. It is important to realize that, for evaluating the Kaplan–Meier estimator, only the events occurring at $\{t_i\}$ are important. That means, between two events, t_i and $t_i + 1$, the survival curve is constant.

2.5 Cox Proportional Hazard (CPH) Regression Model

One of the most popular regression techniques for survival outcomes is Cox Proportional hazards regression analysis. According to Sullivan (2016) the Cox proportional hazards Regression (CPH) model is defined as in Equation (16)

CPH model is used to study the relationship between survival time and covariates in the model. It can also be used to obtain an estimator of the effect size; this estimator takes the form of the hazard ratio which is reported by Exp (B) and 95 percent confidence interval in Table 5.

Assessing the Cox Proportional Hazard (CPH) assumption, the most popular graphical techniques for evaluating the CPH assumption involve comparing estimated In(– In) survival curves over different (combination of) categories of variables being investigated: gender, age, rank at retirement, Length of Service before retirement (LOS) and monthly pension.

$$h(t) = h_0(t)exp\left(b_1X_1 + b_2X_2 + - \cdots + b_pX_p\right) \quad (16)$$

where $h(t)$ is the expected hazard at time t, $h_0(t))$ is the baseline hazard and represents the hazard when all of the predictors (or independent variables) X_1 , X_2 ,. . ., X_p are equal to zero. Notice that the predicted hazard (that is, $h(t)$), or the rate of suffering the event of interest in the next instant, is the product of the baseline hazard ($h_0(t)$) and the exponential function of the linear combination of the predictors. Thus, the predictors have a multiplicative or proportional effect on the predicted hazard (Sullivan, 2016). The hazard function formula for the CPH model, by Kleinbaum and Klein (2016) is defined in Equation (17).

$$h(t,x) = h_0(t)exp\sum_{i=1}^{p}\beta_iX_i \quad (17)$$

where $X_i = (X_1$, X_2 ,. . ., $X_p)$ are explanatory variables, can be converted to a corresponding survival function formula as shown in Equation (18).

$$\hat{S}(t,x) = \mathrm{S_0(t)}\ exp\sum_{i=1}^{p}\beta_iX_i \quad (18)$$

The two quantitative terms considered in any survival analysis, are the survivor function denoted by **S(t),** and the hazard function denoted by **h(t), (**Klein Baum and Klein, 2012**).** The survivor function **S(t),** gives the probability that a person survives longer than some specified time t: that is, **S(t),** gives the probability that the random variable ***T***exceeds the specified time **t**, (**P(T** $>$ **t)**). The survivor function is fundamental to a survival analysis, because obtaining survival probabilities for different values of **t** provides crucial summary information from survival data. The hazard function **h(t),** gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t.

2.6 Assumptions for appropriate use of the CPH model

The fundamental assumption in the CPH model is that the hazards are proportional, which means that the relative hazard remains constant over time with different predictor or covariate levels (Kuitune *et al*. (2021). Time independence of the covariates X_i: the effect of risk factors measured at the beginning of the study period, or at baseline does not change over time (Sullivan, 2016).

According to Emmert-Sterib and Dehmer (2019) there are two significant methods (analytical and graphical) of testing proportional hazard assumption. The underlying idea of both methods is comparison of estimated $\ln[(-\ln)\text{survival curves}]$ and comparison of observed with predicted survival curves. A log-log survival curve is simply a transformation of an estimated survival curve that results from taking the natural log of an estimated survival probability twice.

2.6.1 Analytical Method

Schoenfeld Residual Method is used in testing Cox proportional assumption, a p-value greater than 0.05 does not violate the assumption (Wang *et al*., 2022).

2.6.2 Graphical Methods:

The two graphical methods assess the CPH assumption perform a comparison for each variable one at a time. This means that each covariate is assessed for itself.

2.6.2.1 Graphical Method 1:

In order to understand the first methods, consider the adjusted Survival curves given by Emmert–Sterib and Dehmer (2019) in Equation (19)

$$\hat{S}(t,x) = \mathrm{S}_0(\mathrm{t}) exp \sum_{i=1}^{p} \beta_i X_i \tag{19}$$

Taking In $(-\mathrm{In})$ of Equation (19) gives Equation (20)

$$\mathrm{In}\,(-\mathrm{In}\,\mathrm{S}(\mathrm{t},\mathbf{X})) = \left(\sum_{i=1}^{p} \beta_i X_i\right) + \mathrm{In}\,(-\mathrm{In}\,\mathrm{S0}(\mathrm{t})) \tag{20}$$

Utilizing this expression evaluating two individuals characterized by the specific covariates

$$X_1 = (X_{11}, X_{12}, \ldots, X_{1p}) \tag{21}$$

$$X_2 = (X_{21}, X_{22}, \ldots, X_{2p}) \tag{22}$$

Gives

$$\mathrm{In}\,(-\mathrm{In}\,\mathrm{S}\,(\mathrm{t}, X_1)) - \mathrm{In}\,(-\mathrm{In}\,\mathrm{S}\,(\mathrm{t}, X_2)) = \sum_{i=1}^{p} \beta_i\,(X_{1i} - X_{2i}) \tag{23}$$

From Equation (23), one can deduce that the difference between $\mathrm{In}(-\ln)$ survival curves for two individuals having different covariate values is a constant given by the right hand-side.

For assessing the Cox proportional hazard assumption, one performs such a comparison for each covariate at a time. In case of categorical covariates, all values will be assessed. For continuous covariates, one categorizes them for the comparison. The reason for using Equation (23) for each covariate at a time and not for all at once is that performing such a comparison covariate-by-covariate is more stringent. From Equation (19), it follows that survival curves cannot cross each other if hazards are proportional. Observation of such crosses leads to a clear violation of the Cox proportional hazard assumption. This study employs graphical method 1.

2.6.2.2 Graphical Method 2:

The underlying idea of this approach to compare observed with expected survival curves to assess the Cox proportional hazard assumption is the graphical analog of the goodness-of-fit (GOF) testing. Here, observed survival curves are obtained from stratified estimates of Kaplan -Meier curves. The strata are obtained by the categories of the covariates and the expected

Survival curves are obtained from performing a Cox proportional hazard model with adjusted survival curves, as given by Equation (19).

The comparison is performed as for the $\ln(-\ln)$survival curves, that is, for each covariate one-at a time. For this, the observed and expected survival curve for each stratum is plotted in the same figure for assessment. If for each category of the covariates the observed and expected survival curves are close to each other, the Cox proportional hazard assumption holds.

2.7 Computing the Hazard Ratio

One of the main goals of the Cox PH model is to compare the hazard rates of individuals who have different values for the covariates. Hazard Ratio (HR) is defined as the hazard for one individual divided by the hazard for a different individual. The two individuals being compared can be distinguished by their values for the set of predictors, that is, the $\mathbf{X}'$s. Consider the hazard ratio with p covariates given by Emmert-Sterib and Dehmer (2019) in Equation (19) can be rewritten as in Equation (24)

$$\frac{h(t,X)}{h_o(t)} = exp\sum_{i=1}^{p}\beta_i X_i \quad (24)$$

$$\widehat{HR}_{RSS} = \frac{\hat{h}_i(t|X_i)}{\hat{h}_j(t|X_j)} = \frac{\hat{h}_{0\,(t)}\exp(\hat{\beta}_i X_i)}{\hat{h}_0(t)\,\widehat{\exp}(\hat{\beta}_j X_j)} \quad (25)$$

$$\widehat{HR}_{RSS} = exp\left(\hat{\beta}(X_i - X_j)\right) \quad (26)$$

where $\widehat{HR}_{RSS}$ is the hazard ratio under RSS design, the baseline hazard rate $h_0(t)$ is an unspecified non-negative function of time. It is the time-dependent part of the hazard and corresponds to the hazard rate when all covariate values are equal to zero.

Let $\boldsymbol{X_i} = \boldsymbol{X_j} + 1$, in Equation (26) the hazard ratio reduces to

$$\widehat{HR}_{RSS} = exp(\hat{\beta}) \quad (27)$$

Hence,

$$\hat{\beta} = \log(\widehat{HR}_{RSS}) \quad (28)$$

$\hat{\beta}$, is referred as the log hazard ratio under RSS design. Although the hazard rate $h_x(t)$ is allowed to vary overtime, the hazard ratio is constant; this is the assumption of proportional hazards, which relies on the value of the coefficient of the covariates or its effect on the outcome variable.

Additionally, it is necessary to construct a $(1-\alpha)$ confidence interval for the hazard ratio as in Equation (29)

$$HR(1-\alpha)\% = exp\left(\hat{\beta}(X_i - X_j)\right) \pm Z_{(1-\frac{\alpha}{2})}\hat{Se}\left(\hat{\beta}(X_i - X_j)\right) \quad (29)$$

where,

$$\hat{Se}\left(\hat{\beta}(X_i - X_j)\right) = \sqrt{\widehat{Var}_{RSS}\left(\hat{\beta}(X_i - X_j)\right)}. \quad (30)$$

$$H_0 : \beta_j = 0 \quad (31)$$

The test statistic

$$Z_{RSS} = \frac{\bar{\beta}_j - 0}{Se(\beta_j)} \sim N(0,1) \quad (32)$$

where α Level for this study is 5 percent.

2.8 Adjusted Survival Curves

Survival curves can be obtained adjusted for explanatory variables used as predictors in the Cox model; these are called adjusted survival curves and like Kaplan-Meier curve, this are also plotted as step function. The hazard formula converted before to survival in Equation (18) can be rewritten in Equation (33) as

$$\hat{S}_{RSS}(t,x) = (\mathrm{t}) exp \sum_{i=1}^{p} \beta_i X_i \quad (33)$$

The survival function in Equation (33) is the basis for determining adjusted survival curves under RSS; the covariates are defined as indicated:

X_1 = Sex coded 1 for male and 2 for female.

X_2 = Age at retirement.

X_3 = Rank at retirement (socioeconomic status) defined as categorical variable are coded into four levels: 1 for grade level 1-6, 2 for grade level 7-12, 3 for grade level 13-14 and 4 for grade level 15-17.

X_4 = Length of services in years (LOS)

X_5 = Monthly pension coded as: $1 \leq$ ₦$20,000.00, 2 >$ ₦$20{,}000.00$ but ≤25, 000.00, 3 >₦25, 000.00 but ≤₦50,000.00, 4 >₦50, 000.00 but ≤₦75, 000.00, 5 >₦75, 000.00 but $\leq$ ₦$100,000.00, 6 >$ ₦$100,000.00$ but ≤₦150, 000.00 and 7 >₦150, 000.00.

3. Data Analysis

3.1 Application of RSS in Selection of sample Size

Table 1 shows the application of RSS procedure in selection of 320 sample size from Pension data of Unified Local Government council in Akwa Ibom State used for this study. For cycle one, each set consists of 5 independently selected random numbers and 5^2 formed a complete cycle, each cycle produced 5 judgments ranking order statistic as shown in the last column.

Table1: Ranked Set Sample of Size 320 with K=5, m = 64 for Pensioners retiring from 2016-2020

Cycle	SET I	SET II	SET III	SET IV	SET V	$X_{[1]}$
1	313	278	20	(220)	114	13
	243	166	249	17	95	51
	309	61	40	229	76	81
	(13)	(51)	(81)	165	(222)	220
	274	4	118	45	190	222

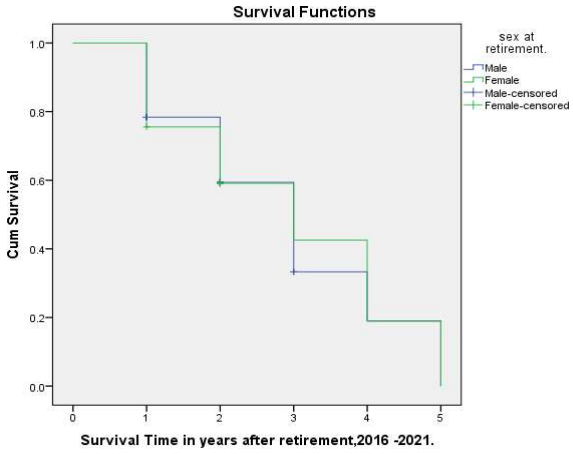
3.2 Result of Fitting Kaplan-Meier Estimator

Using Kaplan- Meier estimator in Equation (15) to evaluate the survival function from the data provides descriptive statistics as shown in Table 2 and the graph of survival function using gender as a factor in Figure 1.

Table 2: Showing descriptive statistics of survival data

Sex at Retirement	Total N	N of Event	Censored	
			Number	Percentage
Male	185	161	24	13.0
Female	135	130	5	3.7
Total	**320**	**291**	**29**	**9.1**

Figure 1: Showing survival function of retirees of Unified Local Government System



3.3 Estimating Survival Rate Using Kaplan-Meier Estimator

At any specific time-interval $(t_i, t_i + 1)$, the survival probability is calculated as the number of subjects surviving $(n_i - d_i)$ divided by the number of subjects at risk (n_i) at t_i. Subject(s) censored within the period are not counted in the denominator for that time interval (Tanujit *et al*. (2020), recall equation (15), Table.3 Shows the summary of estimated survival time for male retirees.

Table 3: Estimation of Survival Rate for Male Retirees using Kaplan-Meier Estimator

Event time (t) in years col.1	Subjects at retirement "lx" col. 2	Died after retirement "d_x" Col.3	Effective no. exposed to risk of dying after retirement "R_x" Col.4	Proportion of dying after retirement "q_x" $\frac{col.3}{col.4}$ Col.5	Proportion surviving after retirement "p_x" (1 − Col.5) Col.6	Cumulative probabilities of surviving after retirement for males Retirees "P_x" Col.7
1	53	13	40	0.3250	0.6750	0.6750
2	41	9	32	0.2813	0.7187	0.4851
3	42	2	40	0.050	0.950	0.4609
4	21	0	21	0.000	1.000	0.4609
5	28	0	28	0.000	1.000	0.4609

Total probability of survival till that time interval is calculated by multiplying all the probabilities of survival at all time intervals preceding that time (Geol *et al*. (2010), by applying law of multiplication of probability to calculate cumulative probability.

From Table 3, five years survival rate for male retirees is given as: $\mathrm{P}(\mathrm{n}) = 0.4609$.

Table 4: Estimation of Survival Rate for Female Retiree using Kaplan-Meier Estimator

Event time (t) in years col.1	Subjects at retirement "lx" col. 2	Died after retirement "d_x" Col.3	Effective no. exposed to risk of dying after retirement "R_x" Col.4	Proportion of dying after retirement "q_x" $\frac{col.3}{col.4}$ Col.5	Proportion surviving after retirement "p_x" (1 __ col.5) Col.6	Cumulative probabilities of surviving after retirement for fe males Retirees "P_x" col.7
1	34	1	33	0.03030	0.9697	0.9697
2	26	4	22	0.1820	0.8180	0.7932
3	21	0	21	0.000	1.000	0.7932
4	30	0	30	0.000	1.000	0.7932
5	24	0	24	0.000	1.000	0.7932

Similarly, from Table 4, five years survival rate for female retirees is given as $\mathrm{P}(\mathrm{n}) = 0.7932$.

3.4 Assessing the Cox Proportional Hazard (CPH) Assumption

Assessing the Cox proportional hazard assumption is a central theme in survival analysis. Statistically, this probability is provided by the survival function

$S(t) = P(T > t)$, where T is a function of time 't'

The most popular graphical techniques for evaluating the CPH assumption involve comparing estimated $\ln(-\ln)$ survival curves over different (combination of) categories of variables being investigated as earlier noted in Section 2.5.

Fitting model for Equation (23) gives the following results, using Cox proportional hazards regression procedure. Figure 2 shows the $\log(-\log)$ survival curve (LML) at different level of monthly pension at retirement.

Figure 2: Showing the log (− log) survival curve at different Level of monthly pension at retirement

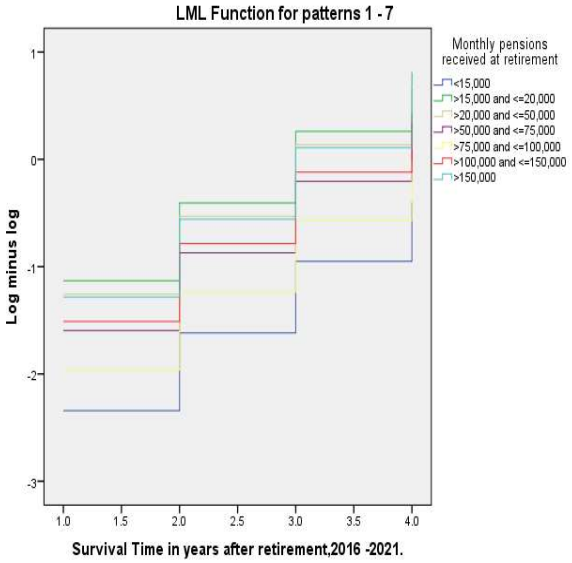


Figure 3 shows the log (− log)survival curve at different level of rank at retirement. Figure 4 shows the log (− log) survival curve at different level of gender at retirement. The assessment is carried out for each covariate at a time and not for all at once. Performing such a comparison covariate-by-covariate is more stringent. Visual inspection of LML plots shows that the survival curves do not cross each other which imply that CPH assumption is not violated.

Figure 3: Showing the log (− log) survival curve at different levels of Ranks at retirement

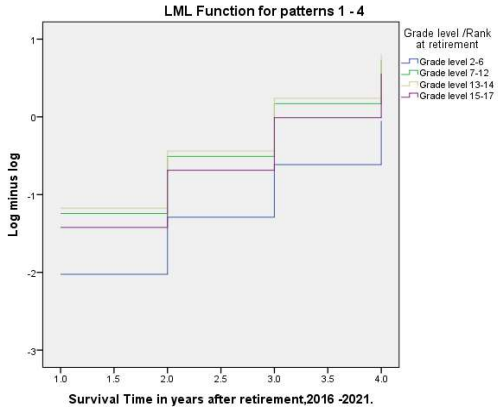


Figure 4: Showing the log (– log) survival curve at different level of Gender at retirement

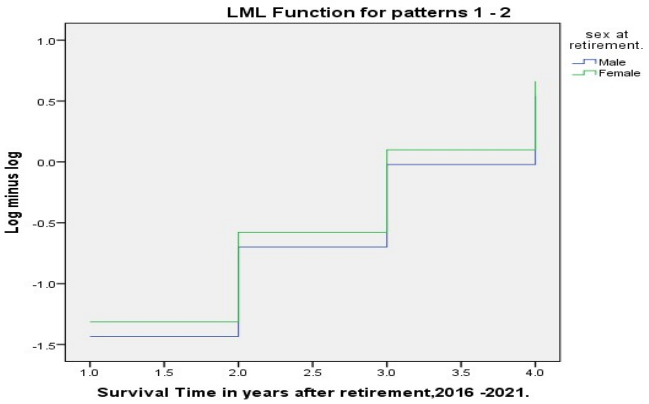


Figure 5: Showing the log (– log) Survival curve by gender and Duration of service before retirement

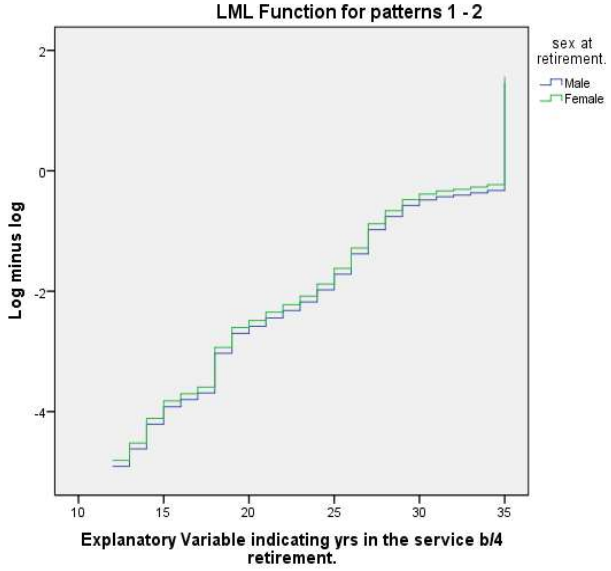
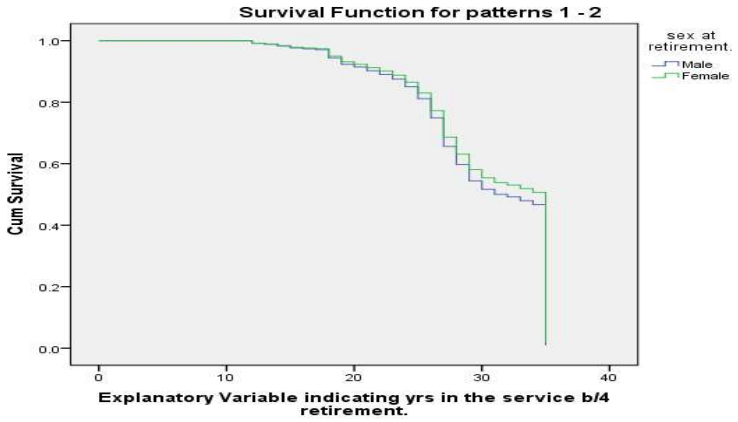


Figure 5 shows another visual assessment of Kaplan – Meier curves, LML plot when event time is the time in years spent in service before retirement controlling for sex, again the curves do not cross each other, meaning that this covariate LOS is time independent variable.

Figure 6: Showing survival distribution of gender by duration of service before retirement



3.4 Assessing Effect of Explanatory variables on Survival time after retirement

Table 5: Showing Variables in the Equation

Covariates	B	SE	Df	Exp(B)	$[1 - \text{Exp}(B)]\%$	95.0% CI for Exp(B)	
						Lower	Upper
Gender	-0.102	0.132	1	0.903	9.7	0.697	1.171
Age at retirement	0. 070	0.028	1	1.073	-7.3	1.015	1.134
Rank at retirement:			3				
Status (1)	0.246	0.563	1	1.279	-28.0	0.259	2.355
Status (2)	0.006	0.439	1	1.006	-0.6	0.425	2.380
Status (3)	-0.158	0.216	1	0.854	14.62	0.560	1.304
LOS	-0.015	0.011	1	0.985	1.5	0.964	1.007
Monthly Pensions	-0.012	0.107	1	0.988	1.2	0.802	1.218

4. Results and Discussion

RSS has been satisfactorily used in selection of sample size for this study as seen in Table 1, without loss of generality, it is assumed that there is perfect ranking that is, within each cycle one unit belongs to each rank order statistics and hence *m* ranked units from each circle are included in the measurement.

Fitting the model in Equation (15) provides descriptive statistic in Table 2. Out of 320 sample size, male and female retirees are 185 and 135 respectively. Out of 291 events, 161 male and 130 female retirees experienced event of interest respectively. The 29 subjects censored, 24 are male while five are female, those that could not survive after retirement.

From the analysis, 13 males died after first year of retirement, 9 died after two years and two died after three years of retirement respectively. Five females died after the first two years of retirement as summarizes in Tables 3 and 4 respectively. It can be inferred that any pensioner that survived the first three years after retirement has higher probability of surviving longer after retirement.

Survival function can be determined using the Kaplan–Meier curve, survival rates indicate the number of patients in whom no event has occurred up to a certain point in time (Zwiener *et al*., (2011). In non-medical field this should be interpreted with care, because in medical field

the event of interest is always time to death. In this study survival rates indicate the number of pensioners whose event of interest did occurred to a certain point in time. When stating survival rates, it is important to also state the point in time to which it corresponds. From Tables 3 and 4, the overall probability of surviving until at least the end of five years of study; for male estimated survival rates is about 0.4609, and for female estimated survival rate is about 0.7918. This can be interpreted as follows: five years after retirement, we can expect 46 percent male and 79 female pensioners to be alive respectively.

Again, the assumption of the Cox Proportional Hazard (CPH) was assessed Visual inspection of Figures 2, 3, 4, and 5 respectively shows that the LML survival curves do not cross each other meaning that assumption of proportional hazard is not violated.

Thus, assessing the effects of five covariates on the survival time, four variables demonstrated an effect on overall survival; they are gender, rank at retirement, monthly pension and Length of Service (LOS).

Age-at-retirement is continuous variable, the coefficient of this variable is 0.070, Exp (0.070) = 1.073 is hazard ratio and is greater than one, the 95 percent confident interval for the hazard ratio is [1 .015, 1.134]. This variable has 1.073 as hazard ratio, therefore there is about 7.3 percent decreased in survival rate to a one-year increase in age (or the expected hazard is 1.07 times higher in a person who is one year older than another), holding other predictors constant. Age is potential decrease of survive time. Age is not a significant determinant of survival time with hazard ratio 1.073 (HR > 1). This study has shown that some pensioners died between the ages of 51 to 60 years in the first three years after retirement, so age does not improve survival time.

For gender in Table 5, the estimated coefficient is -0.102 and $\text{Exp}(-0.102) = 0.903$, which is the hazard ratio and is less than one. The 95 percent confidence interval for this hazard ratio is $[0.697, 1.171]$. Sex is categorical variable, male as reference group; this means that female's survival time after retirement is increased by 9.7 percent more as compare to the males' counterpart. Sex is very strong time independent covariate which does not change over time. Mwakala (2013), in her work noted that there is significant difference between male and female survival time and concluded that gender does indeed determine the rate of survival after retirement in Kenya, in fact that male are about 2.1 times more likely to die compare to their female counterparts. This study has demonstrated that sex also influence survival time with hazard ratio less than one, female having about 0.79 survival rates and male about 0.46 survival rates, this has shown that the result is consistence in the existing result that female tend to live longer than their male counterpart. [see Mwakala (2013)].

Rank at retirement determine the salary grade level attained at retirement and is coded into four categories, grade level 2-6, 7-12, 13-14 and 15-17 respectively, grade level 15-17 served as referenced group. Grade level 2-6 with estimated coefficient of 0.246, Exp (0.246) = 1.279, this hazard ratio is greater than one, the 95 percent confidence interval is [0.259, 2.355], that is to say the pensioners in grade level 2-6 are about 28 percent decreased in survival rate compare to those in grade level 15-17. Grade level 7-12 has estimated coefficient 0.006 with hazard ratio 1.006, hazard ratio greater than one (HR>1), the 95 percent confidence interval for hazard ratio is [0.425, 2.380], this implies that pensioners in grade level 7-12 are about 0.6 percent decreased in survival rate after retirement compare to those in grade level 15-17. Again, grade level 13-14 has estimated coefficient -0.158, Exp (-0.158) gives the hazard ratio as 0.854 (HR<1) which is a good prognosis for the outcome variable, the 95 percent confidence interval is [0.560, 1.304], that is to say pensioners in grade level 13-14 are about 14.6 percent increase in survival rate after retirement to compare to pensioners in grade level 15-17. Rank at retirement, some researchers termed it socio-economic status is well-known to be a good predictor of mortality (Sullivan, 2016), people in higher socio-economic have lower mortality

at any given age and longer average survival than those in lower grade level. This was seen in Table 5 as those in grade level 2-6 and 7-12 tend to have higher hazard ratio compare to those in grade level 13-14.

There are disparities in monthly pensions earned by pensioners resulted from different grade level, the estimated coefficient of this variable is –0.012, the hazard ratio is 0.988 (HR<1) which is a good prognosis for the outcome variable, the 95 percent confidence interval is [0.802, 1.218]. Thus, monthly pension increased survival rate by about 1.2 percent.

Mwakala (2013) mentioned that there was a significant difference in probability of surviving among the retirees based on monthly pension received as she relates socio-economic status with monthly pension. In this study, monthly pension is seen to improve survival time. Apart from monthly income, other socio-economic variables like occupation of the pensioner after retirement, academic attainment while in service and family size should be made available by the retirees to be accounted for in subsequent study.

Length of service before retirement (LOS) has a good prognosis for the outcome variable. This covariate has estimated coefficient of –0.015, Exp (- 0.015) gives hazard ratio of 0.985, (HR<1) and the 95 percent confidence interval for hazard ratio lies between [0.961, 1.006]. This variable is very important to any retiree; pension and gratuity are based on the years spent in service before retirement. This variable improves survival time after retirement by about 1.5 percent. Length of Service (LOS) is years spent in service before retirement, this covariate has hazard ratio less than one invariable infer to improve survival time. Possibly, pensioners while in the service might have use this time to prepare for their exit from service, knowing that retirement is a different phase of life that requires strategic preparation without which is difficult to survive in the first three years. Figure 6 shows survival distribution of gender by duration of service before retirement adjusted for gender, this distribution revealed that retirees spent between 12-35 years before retirement. In addition, terminal benefits depend on years of qualifying service (length of service). Though in the work of Ajayi *et al*. (2014) this covariate, duration of service before retirement does not have effect on longevity of retired Academic Staff of University; this may depend on the definition of event of interest, presence of other covariates in the model and statistical tool used. They used Analysis of Variance (ANOVA) to estimate the effect of predictors in the model whereas this study made use of CPH regression model which gives a better result than ANOVA.

5. Conclusion

In view of the above analysis, it is observed that RSS yields a good representative samples, its asymptotic property is seen in the Kaplan-Meier estimate's step function with jumps at event times, the ties in the event time is grouped into set event time. KM estimator under RSS design estimated survival rates for male is about 0.46 and female 0.79 in Unified Local Government Council of Akwa Ibom State. The Log rank test revealed no significant difference between male and female survival function with (p–value > 0.05).

As seen in the work of Chandra *et al*. (2018), RSS is cost–effective, time saving and precise method of sample selections. It provides a better estimate of the characteristic under study. It yields a good representative samples, RSS contains information across the entire population

Cox Proportional Hazard assumption was not violated, the covariates are all time independence variables. This study shows that gender, rank at retirement, Length of Service (LOS) and monthly pensions are strong determinants of survival time after retirement.

With this question what make one pensioner live longer compare to another pensioner, it is suggested that study should be done to ascertain causes of early death among pensioners for possible interventions by the government.

References

[1]. Ajayi, M., Kehinde, S., Lucky, M. and Thaga, K. (2014). Effect of Post Retirement Occupation on Survival of Academic Staff Retirees: A Case Study of University of Ibadan Nigeria. www.isorjournal.org

[2]. Al-Omari, A. and Bouza, C. (2014). Review of ranked set sampling: Modification and Application, 3(3), 215-235. https/www.research.net/publication

[3]. Bamia, C., Trichopoulou, A. and Trichopoulou, D. (2008). Age at Retirement and Mortality in General Population Sample the Greek Epic Study. *American Journal of Epidemiology*, 167(5),561-569.

[4]. Chen, W., Xie, M. and Wu, M. (2013). Parametric estimation for the scale parameter for scale distribution using moving extremes ranked set sampling. Statistics and *Probability letters*, 83(9), 2060-2065. Doi: 1 0. 1 0 1 6 1 /. S p l. 2 01 3 .0 5.0 1 5, https:// www.research.net/publication/26828046.

[5]. Chandra, G., Pandey, R., Bhoj, D., Nautiyal, R., Ashraf, J., and Verma, M. (2018). Ranked set approach for estimating response of developmental programs with linear impacts under successive phases. *Pakistan Journal of Statistics*, 34(2), 163-176.

[6]. Daepp, M., Hamilton, M., West, G., and Bettencourt, L.(2015). The mortality of Companies. *Interface.* https://www.doi.org/10,1098/rsifi2015.0120

[7]. Emmert – Sterib, F. and Dehmer, M. (2019). Introduction to survival analysis in practice. *Machine learning knowledge Extraction,* 1013-1038

[8]. Fields, R., Busam, K., Chou J., Pangeas, K., Pulitzer, M., Kraus, D., Brady, M and coit, D. (2011). Recurrence and survival in patients from a single institution. *Annal Surgical Oncology*, 18.

[9]. Goel, M., Khanna, P. and Kishore, J. (2010). Understanding Survival analysis: Kaplan-Meier estimate. *Internarional Journal of Ajurveda Research.* 1(4), 274-278. Wolters Kluwer-MedknowPublication.doi:10.4103/0974-7788.76794 https://www.ncbi.nlm.gov>

[10]. Gaur, A. Mahajan, K. and Arora, S. (2013). A non-parametric test for a multi-sample scale problem using ranked set data. *Statistical Methodology*, 10, 85-92.

[11]. Kleinbaum, G. and Klien, M. (2012). Survival Analysis a self-learning text. *Third edition 10-174, Springer.*

[12]. Khan, L. and Shabbir, J. (2016). Improved ratio-type estimator in ranked set sampling using two concomitant variables. *Pakistan Journal of Statistics and Operational Research,* 12(3), 507-518.

[13]. Kuitune, I. Ponkilainen, V., Uimonen, M., Eskehinen, A, and Reito, A. (2021). Testing the proportional hazards assumption in Cox regression and dealing with possible non-proportionality in total joint arthroplasty research methodological perspective and review. *BMC Musculoskeleta lDisorders,* 22489. https://doi.org/10.1186/s12891-021-04379-2

[14]. Lesem, M., Tran-Johnson, T., Riesenberg, R., Fiefel, D., Anen, M., Fishermen, R. Spyker, D., Kehne, J., and Cassella, J., (2011). Rapid acute treatment of agitation in individual with schizophrenia: Multicenter, randomized, placebo-controlled study of inhaled loxapine.

[15]. Min, Y., Zhang, G., long, R., Anderson, T., and Ohland, M. (2011). Non parametric survival analysis of the loss rate of undergraduate engineering students. *Journal of English Education*, 349-373.

[16]. Mwakala, M. (2013). Modeling time to death for retirees in Kenya. MSc dissertation. University of Nairobi. https://.acke>bitstream>hande pdf (Retrieved on 18/11/2020)

[17]. Richardeau, F. and Pham, T. (2012). Reliability of multilevel converters: Theory and applications. *IEEE Trans Ind. Electron*, 60, 4225-4233.

[18]. Sullivan, L. (2016). Survival Analysis: Cox proportional hazards regression analysis Boston University School of Public Health. https://sphweb

[19]. Saxen. (2018). Determine sample size. *Taxton State University*. https://www.tarleton.

[20]. Tanujt, D., Anish, M. and Sounak, C.(2020). A practical overview and reporting strategies for statistical analysis of survival studies. *Chest Journal*, 158(1), 539-548, https://doi.org/10.1016/j.chest.2020.03.015

[21]. Wolfe, D. (2012). Ranked Set Sampling: Its Relevance and Statistical Inference. *Journal of International Scholarly Research Network Probability and Statistics*, 1-32. doi: 105402/2012/568385.

[22]. Wang, Z., Zhang, L., Xu, F., Ham, D. and Lyu, J. (2022). The association continuous renal replacement therapy as treatment for sepsis associated acute kidney injury and trend of lactate trajectory as risk factor of 28-days mortality in intensive care units. *BMC Emergency medicine,* 22,32 h t tp : / / d o i . o r g / 1 0 . 1 186/s/2873-022-005896

[23]. Zwiener, I., Blettner, M. and Hommel, G. (2011). Survival Analysis. *Deutsch Arzteb International*, 108(10), 163-169. Doi 10.3238/arzteb//2011.0163

[24]. Zhang, Z. (2016). Statistical description for survival data. *Annuals of Translational Medicine*, 4(20).