IRAQI Academic Scientific Journals

KHAZAYIN OF ECONOMIC AND ADMINISTRATIVE SCIENCES (2025) (02) (03): P(187-198)

خزائن للعلوم الاقتصادية والإدارية

KHAZAYIN OF ECONOMIC AND ADMINISTRATIVE SCIENCES

ISSN: 2960-1363

Adjusted Kaplan-Meier Survival Estimator Ranks for a Set of Medical Data

Alaa H. Jalob [1], Hiader K. Raheem[2] , [3]Muhammed J. Kadhim

[1]Department of Studies and Planning / AL- Karkh University of Science, Baghdad, Iraq
[2] Department of Administrative and Financial Affairs/ AL- Karkh University of Science, Baghdad, Iraq
[3]Department of Power Mechanical Techniques, Institute of Technology, Middle Technical University, Baghdad, Iraq
[1]alaahasaan@kus.edu.iq , [2]haider@kus.edu.iq , [3]muhammedkadhim74@gmail.com

Abstract: This study presents four methods for the purpose of adjusting the estimation ranks of the Kaplan-Meier survival function, wconsidered alternative non-parametric estimates of the grouped censored data, thus obtaining estimates that are the rank-adjusted estimator , the Smoothing survival curves estimator, the Smith-Waterman estimator and the histogram estimator, in the applied aspect a simulation system is designed for the purpose of generating controlled aggregated data of three sizes (50 , 80, 100) for the purpose of calculating the estimate of the survival function for three sample sizes and comparing with the Kaplan-Meyer estimate, which serves as a reference model, and then calculating performance measures where it was found that the graph estimator comes first for all sample sizes followed by the Smoothing survival curves estimator by increasing the sample size..

Keywords: Kaplan-Meier survival estimation, Censored data, Rank-Adjusted estimator, Smoothing survival curves, Smith-Waterman, Histogram estimators.

DOI: 10.69938/Keas.25020313

1.Introduction:

Survival analysis is one of the most important branches of statistics that is used in the analysis of medical data. The main goal of survival analysis is to estimate the probability of survival and interpret it at a specific point. Medical studies, clinical trials, life science, epidemiology and diseases in which control and incomplete data appear frequently depend on the field of survival analysis that plays a pivotal and important role in it (Collett, 2023; etikan, Abubakar, and Al-Qasim, 2017; Joel, Khanna, and kashour, 2010). Censorship occurs when the full time of an event (e.g. death or illness) is not fully observed for all cases, resulting in partial information about survival times. In such trials or settings, accurately estimating survival function is challenging for clinical decision-making, evaluating treatment effectiveness, and development (Collett, 2023).

The Kaplan-Meier method or estimate, which is often denoted by (K-M), is still the most common and widely used non-standard estimate of the survival function given by (Kaplan-Meier, 1958) because it has simplicity, interpretability and strength under grouped censored data (Itkan et al., 2017; Joel et al., 2010). The most common concept about the definition of (K-M) is the conditional probability of survival during certain times of events, which for its frequent use and features is considered to be the standard model or reference model in estimating the survival function, where it assumes that lifetimes are continuous through time and does not take into account the real reality

and what is encountered during the Applied side during practical practices (nematolahi et al., 2020). Due to the frequent recording of medical data or undergoing interval monitoring and periodic follow-ups during applied practices, which determines or impairs the efficiency of the method (Collett, 2023).

From the foregoing, there was a need to propose alternative non-parametric estimates that can deal with grouped censored data that are collected periodically or take into account periodic monitoring according to the established routine or administrative protocols during the compilation of monitored survival data to reduce the amount of bias in sample sizes and thus provide clear and smooth estimates of the survival function (H. okerdl, 2012; Maron and H. okerdl, 1986). This study modifies the Kaplan-Meier estimation ranks using four methods : (I) rank-adjusted estimator (Les and Schumann, 2010; kunitomo and Matsushita, 2008), (II) Smoothing survival curves estimator (III) The Smith-Waterman estimator (Chen, Ferris, and Turk, 2008; comet et al., 1999; Khajeh said, Paul, and Piru, 2010; Mott, 1992), and (IV) the graph estimator (Friedman and diakonis, 1981; Chan and eroldi, 2014; Kontkanen and milim Oshki, 2007; Chen and Kelton, 2001). For the purpose of addressing the limitations imposed by practical reality on the Kaplan-Meier estimate that contradicts the assumption of continuous observation time by making adjustments to the contraction of the denominator or adjustment to the intervals during data aggregation or exponential decay.

This study uses a rigorous simulation framework to generate synthetic survival data under exponential distributions of both event and control times. By applying each estimator to data sets of different sample sizes (N = 50, 80, 100), we evaluate and compare their performance relative to the Kaplan-Meier estimator using multiple statistical measures. These metrics include mean squared error, mean absolute error, square root error, bias, variance, mean absolute percentage error, and R2. Through these comprehensive comparisons, the study aims to determine whether any of the proposed estimators offers superior or comparable performance in estimating the survival function under combined control conditions.

In the applied aspect, the method of generating controlled medical data was used based on the design of a simulation system that generates exponential distributions for each of the sample sizes from the Times of events and control, then applies and calculates the formula of each of the four estimates in addition to the reference estimate (Kaplan-Meier) for different sample sizes (N = 50, 80, 100), then we evaluate and compare the performance of each formula of the four methods with each other – Meyer is a reference model, these metrics include mean squared error, mean absolute error, square root error, bias, variance, mean absolute percentage error, and R2, so it is easily possible to determine which of the methods of adjusting the ranks or weights of the Kaplan-Meier estimate is the best under practical practices facing continuous-time assumptions, especially in contexts where sample sizes are rather small or medium.

2.Literature Review and Hypotheses

Kaplan-Meier estimation is one of the nonparametric methods used to estimate survival functions for medical, clinical and epidemiological research (Itkan, Abubakar, & Al-Qasim, 2017; Joel, Khanna, & Kashour, 2010). Which were identified by researchers to estimate the survival probabilities of patients in clinical studies over time and during the presence of censored data. Several studies have shown that the Kaplan-Meier standard estimate often suffers from limitations when adjusting the rank to improve accuracy, especially when dealing with small sample sizes or a heterogeneous data set (Morris et al., 2019; nematolahi et al., 2020).

Many researchers have proposed rank adjustment for survival abilities based on addressing biases and inefficiencies, where he (mantel, 1966) was the first to introduce the rank adjustment convention that means rank-dependent statistics in survival analysis and then laid the foundation for further improvements, more recently he (Lise & Schumann, 2010) conducted a modified rank test focused on the detection of interaction effects, which are considered nonparametric methods that are not directly applied to Kaplan-Meier estimates but the principle of rank conversion in these tests focus on the benefits of rank adjustment to improve survival estimates.

In this context of rank adjustment, several methods of rank adjustment have been proposed. (Jung, 2008) introduced a weighted-rank statistical method designed for paired survival data, which is

useful in calculating the reliability between sightings, also (nematolahi et al., 2020) propose and present an estimation method based on improving the Kaplan-Meier estimate based on taking each time a partially ordered set (Pros), which has been shown to enhance the efficiency of the estimator under the specified sampling conditions.
In addition to the above, (Friedman&diakonis, 1981) and (etcherdel, 2012) have shown that graph – based estimation of the survival function requires non-standard smoothing for density estimation to achieve a consistent and unbiased estimate, as (kontkanin &milim oschke, 2007) showed that it is possible to balance the complexity of the model and its accuracy through a graph based on the minimum description length which is convertible to Kaplan-Meier estimates adjusted for rank.
The above reference review prompted us to present four methods or methods for rank adjustment to estimate the Kaplan-Meier survival function, which is a reference model, and then make a comprehensive comparison between these estimates, which are the rank-adjusted estimator (Les & Schumann, 2010), the Smoothing survival curves estimator (nematolahi et al., 2020), the Smith-Waterman estimator (Jung, 2008), and the graph estimator (kontkanin&milim oschke, 2007) relying on performance measures to choose the best method that helps improve or adjust the rank for the Kaplan-Meier estimate helps to balance efficiency, accuracy, robustness and practicality of practices where accurate estimates are important in clinical decisions (Collett, 2023).

3.Research Methodology

This paper use methodological form grounded in nonparametric survival function analysis. The main objective is to estimate the survival function as formula $S(t) = Pr(T > t)$, where T is a non-negative random variable representing time-to-event, using censored data. In the with right-censoring, the observed data consist of pairs (Y_i, δ_i), where $Y_i = \min(T_i, C_i)$ is the observed time, C_i is the censoring time, and $\delta_i = I(T_i \leq C_i)$ is the event indicator (Nieto & Coresh, 1996). given a sample of size n, our objective is to estimate $S(t)$ is nonparametrically ، The Kaplan-Meier serves as the reference or benchmark model (Etikan et al., 2017). Let we ordered observed times as $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$, hence let $t_1 < t_2 < \cdots < t_k$ represent distinct event times. At each event time t_j, d_j represents the number of events and n_j the number at risk. The Kaplan-Meier estimator is computed from :

$$\hat{S}_{KM}(t_j) = \prod_{t_i \leq t_j} (1 - d_i/n_i) \qquad (1)$$

Here (Leys & Schumann, 2010) , we present four methods to adjust the rank of the Kaplan-Meier estimate of the survival function for three sample sizes generated using the simulation system according to the method of observation data Rank-Adjusted, Smoothing survival curves, Smith-Waterman, and Histogram estimators, which are mathematically adjusted or modified based on adjusting the rank of the estimated Kaplan-Meier survival function, and then choosing the best method from the above methods to determine it in adjusting the rank of the survival data set based on performance measures to reduce or control the inconsistency that guides the basis of the assumption underlying the estimation of the Kaplan-Meier survival function (Goel et al., 2010), which assumes that the survival time is continuous while not taking into account periodic monitoring, protocols or inconsistencies during the combined monitoring.

4. Methods

4.1.Rank-Adjusted Estimator

The estimate of the Rank-Adjusted survival function is a non-parametric estimate based on the subtraction of a constant (usually C) from the Kaplan-Meyer denominator for the purpose of calculating the amount of bias of a small sample, when this constant is subtracted from the denominator (Kunitomo & Matsushita, 2008), we enhance the accuracy when calculating the survival estimate from the collected or highly controlled data and also prevent duplicate risks in estimating this function $Y_i = \min(T_i, C_i), \delta_i = I(T_i \leq C_i)$, where T is a random variable as time-to-event.

Where the censoring time is C_i , $\delta_i \in \{0,1\}$ We observe data $\{(Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)\}$ and use Kaplan-Meier Estimator (Reference Model) , Order distinct event times (Nematolahi et al., 2020) :

$$t_1 < t_2 < \cdots < t_k$$

Upward bias in $\hat{S}(t)$ It is used Jackknife bias reduction idea (Efron ,1967) Adjust denominator with use constant C as $C = \frac{d_j}{2}$ then we get $n_j^* = n_j - \frac{d_j}{2}$,Where **C** is a small positive constant for each event time t_j , (Leys & Schumann, 2010) :

$$q_j = \frac{d_j}{n_j^*} = \frac{d_j}{n_j - \frac{d_j}{2}} \qquad (2)$$

By using Equ.2 , Then a formula have :

$$\hat{S}_{RA}(t) = \prod_{t_j \le t} (1 - q_j) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j - \frac{d_j}{2}}\right) \qquad (3)$$

4.2. Smoothing survival curves Estimator

This means the Rank Adjustment of the estimated survival function, that the Smoothing survival curves estimate divides the survival time into separate periods such as intervals (bins) ((Wu & Kolassa, 2024), thus smoothers the survival curve, thereby reducing the variance in order to reduce the amount of bias by adjusting the order, that interval grouping leads to a decrease in variance (Kim et al., 2003) , divides the timeline into periods (bins):

$$I_k = [b_k, b_{k+1}) \qquad (4)$$

Can be defined as b_k and b_{k+1} are the interval boundaries of I_k , and the $k = 1,2,\ldots,m$ is number of intervals (bins) to determine m (number of intervals) using Sturges' Rule:

$$m = \lceil \log_2(n) + 1 \rceil$$

where n is the sample size , for each interval in Equ.4, The quantities are calculated n_k the number of patients initially at risk at I_k, this is the count of individuals whose observed time $Y_i \ge b_k$ (i.e., they are still under observation just before b_k).

$$d_k = \text{Number of failures (events) that occurred within } I_k$$

This is d_k the count of individuals whose event time $Y_i \in [b_k, b_{k+1})$ and $\delta_i = 1$ (i.e., the event occurred within this interval and it was not censored) , we apply the adjustment rank to factors (0.5) as constant c , similar to event probability within I_k the interval in Equ.4 (Jung, 2008) :

$$q_k = \frac{d_k}{n_k - 0.5} \qquad (5)$$

Bay using q_k as corrected factors in Equ.5 to determined Survival estimate up to interval k:

$$\hat{S}_L(I_k) = \prod_{l=1}^{k} (1 - q_l) = \prod_{l=1}^{k} \left(1 - \frac{d_l}{n_l - 0.5}\right) = \prod_{I_l \le t} \left(1 - \frac{d_l}{n_l - 0.5}\right) \qquad (6)$$

In Equ.6 include all intervals I_l that is end $b_{l+1} \le t$.

4.3. Smith-Waterman Estimator

The Smith-Waterman Estimator is a nonparametric survival estimator that applies sequence analysis-based local alignment notions to survival functions (Chen et al., 2008). It improves robustness to irregular event distributions by focusing local survival regions. It computes survival probabilities using weighted local risk sets, addressing the bias-variance trade-offs in both sparse and heterogeneous censored datasets. T is a time-to-event random variable. Observed data using

censored data $\{(Y_i, \delta_i)\}_{i=1}^{n}$ Arrange observed times and distinct event times (failures) For each event time t_j, define a local neighbourhood (window), (Mott, 1992) is :

$$W_j = [t_j - h, t_j + h] \quad (7)$$

Can be defined h as the bandwidth or also called the window size , which determines the local area for the time of the event t_j, thus calculating the set of possible risks and events during the time for each window in Equ.7 , (Härdle, 2012) as :

$$n_j^{(W)} = \sum_{i=1}^{n} I\left(Y_i \in W_j\right)$$

The number of events in window W_j :

$$d_j^{(W)} = \sum_{i=1}^{n} I\left(Y_i \in W_j, \delta_i = 1\right)$$

Now, the probability of a local event can be calculated (adjusted for window) :

$$q_j^{(SW)} = \frac{d_j^{(W)}}{n_j^{(W)} - 0.5} \quad (8)$$

The constant 0.5 is adjustment to bias reduction, and define Smith-Waterman Estimator using Equ.8 as estimated survival function is:

$$\hat{S}_{SW}(t) = \prod_{t_j \le t} \left(1 - q_j^{(SW)}\right) = \prod_{t_j \le t} \left(1 - \frac{d_j^{(W)}}{n_j^{(W)} - 0.5}\right) \quad (9)$$

4.4. Histogram Estimator

The Histogram Estimator can be defined as a non-parametric approach designed for the purpose of presenting a survival function that by taking advantage of the behaviour of grouped (binned) data, as a result, the survival function is calculated as an Empirical Cumulative Distribution Function (ECDF) over predefined periods (Härdle, 2012), since the available data is grouped or interval-censored data, this estimator gives a smoother and more stable estimate of the survival function through projections of events within boxes of equal width (fixed bandwidth) all events are separate in Kaplan-Meyer because he processed the notes Separately censored data by density estimates within each container and also the summation of contributions, therefore, this method mitigates the effect of local fluctuations and sharp jumps in the estimated curve, which is useful in smaller samples, to partition the time into m , as $I_k = [b_k, b_{k+1})$ when $k = 1,2, \dots, m$ and $b_1 < b_2 < \cdots < b_{m+1}$, and bin width:

$$\Delta_k = b_{k+1} - b_k \quad (10)$$

Now we can have obtained the observed events in each bin, (Kontkanen & Myllymäki, 2007) :

$$c_k = \#\{i : Y_i \in I_k,\ \delta_i = 1\} \quad (11)$$

From Equ.10 and Equ., we get an estimate of the probability density function in each bin k as :

$$\hat{f}_k = \frac{c_k}{n \cdot \Delta_k}$$

The estimated Empirical Cumulative Distribution Function (ECDF) at time t is the sum of $\hat{f}_k$:

$$\hat{F}(t) = \sum_{b_k < t} \hat{f}_k \cdot \Delta_k \quad (12)$$

From (Freedman & Diaconis, 1981) , we substituting Equ.10 and $\hat{f}_k$:

$$\hat{F}(t) = \sum_{b_k < t} \frac{c_k}{n \cdot \Delta_k} \cdot \Delta_k = \sum_{b_k < t} \frac{c_k}{n}$$

Now the formula of estimate the survival function, by (Marron & Härdle, 1986) have :

$$\hat{S}_H(t) = 1 - \sum_{b_k < t} \frac{c_k}{n} = \frac{n - \sum_{b_k < t} c_k}{n} \quad (13)$$

5. Simulation Design and Medical Dataset Generation

In The simulation part will be presented in this part of the study, by designing simulations to generate experimental medical survival data under different control. The basis is to evaluate the performance of various survival estimates under realistic medical conditions , Survival times were simulated based on the Weibull distribution with the figure parameter =1.5 and the scale parameter =12, which gives a picture of the survival patterns observed in breast cancer patients as reported in (Allen et al., 2009),the survival function is derived as follows:

$$S(t) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\kappa}\right) \qquad (14)$$

The independent control time was assumed based on a uniform distribution $C_i \sim U(5,20)$,which represents the follow-up periods in typical clinical studies. Then calculate the observed time for each subject as follows:

$$T_i = \min(S_i, C_i), \delta_i = I(S_i \leq C_i)$$

where δ_i is the event indicator (1 = event occurred, 0 = censored).

Three different sample sizes (N = 50, 80, 100) were used to find out the effect of sample size on the performance of estimates . The Kaplan-Meyer estimator has been used as a reference model due to its wide application in survival analysis. It was compared with four alternative estimators: the rank-adjusted estimator, the Smoothing survival curves estimator, the Smith-Waterman estimator, and the graph-based estimator. These estimators represent modifications of the estimation of the standard survival function, these simulations are designed using Python programming language algorithms for the purpose of simulating realistic medical survival scenarios.

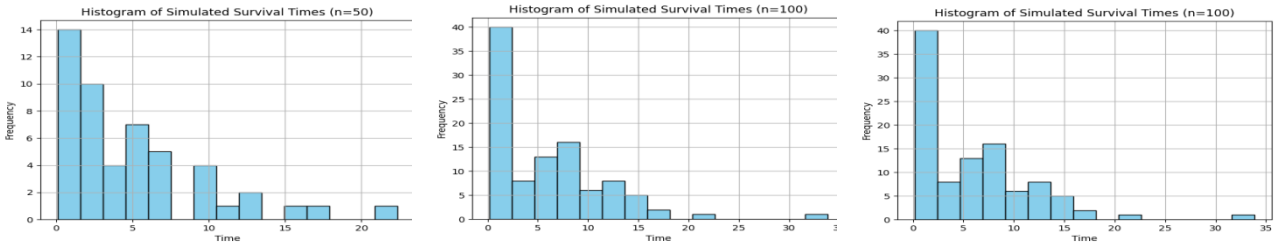

Figure 1: Represents The Histogram For Generating Data For Various Sample sizes

6.Discussion of Results

6.1. Overview Results

The results of calculating the four estimates will be presented, which are the rank-adjusted estimator, the Smoothing survival curves estimator, the Smith-Waterman estimator and the Histogram Estimator compared with the true value of the survival function, which is the Kaplan-Meier estimate, which was assumed as a reference model whose rank was adjusted by four methods. Here in this part of the study, visual comparison according to the graph and curve is very important for the purpose of knowing the behavior and correspondence of the estimates of the survival function.

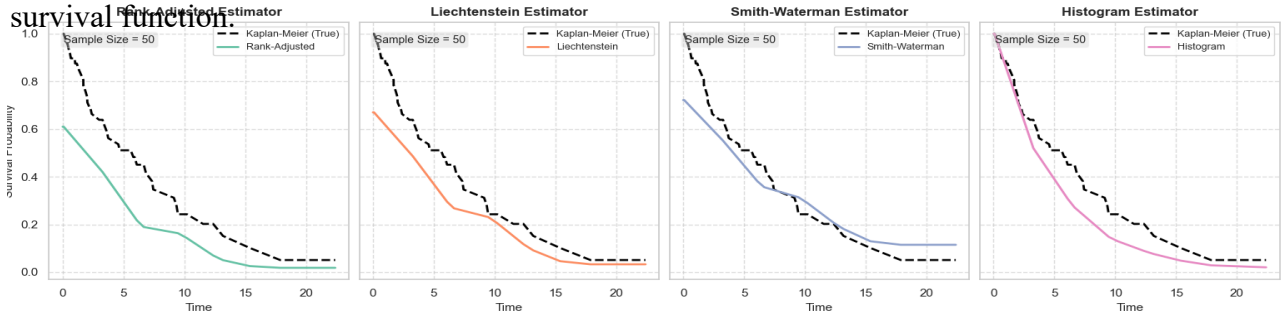

Figure 2: Represents the Comparison Curve of Benchmark Model with the Others Estimators at n=50

We note from the above Figure that the Histogram method of adjusting the rank of the Kaplan-Meier survival function was closer than the other three estimates, because the rank-adjusted

estimate moved significantly away from the true value, followed by the Smoothing survival curves estimate, with Smith-Waterman approaching significantly in the average values while decreasing by a lower level, finally it turned out that the Histogram estimate gives a clearer proximity of all the true values at the top and bottom to estimate the survival function with a noticeable deviation in the curve.

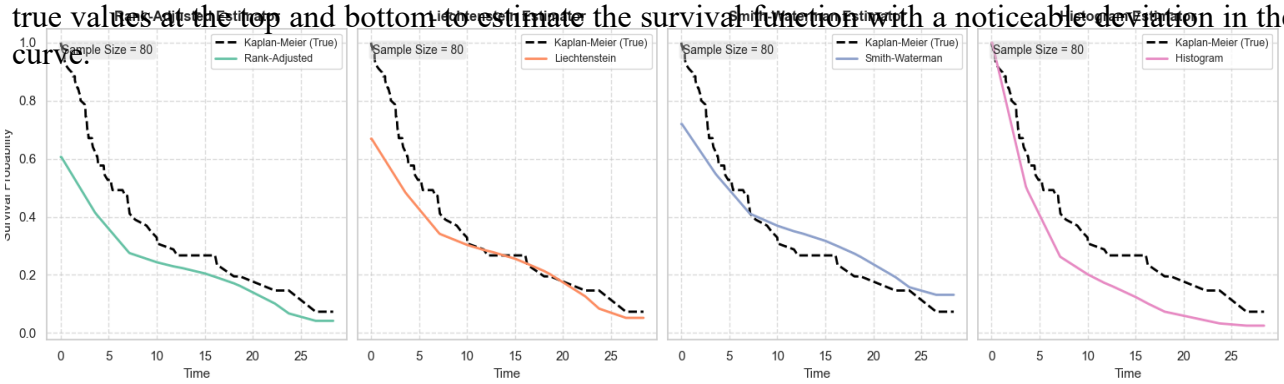

Figure 3: Represents the Comparison Curve of Benchmark Model with the Others Estimators at n= 80

Figure 2 shows a comparison of the Kaplan-Meier survival function using four proposed methods for adjusting the rank of survival times, we note that the rank adjustment method moved away a lot and began to be consistent with the Kaplan-Meier curve estimate for Time 20 and above (the tail of the curve) while it turned out that the Smoothing survival curves estimate followed the behavior of the Kaplan-Meier curve gradually at the tail of the curve, at the other end of the figure, we are shown the estimate of the graph, which significantly follows the curve of the true value of the model A reference that is considered more and gives a more impression of follow-up and consistency.

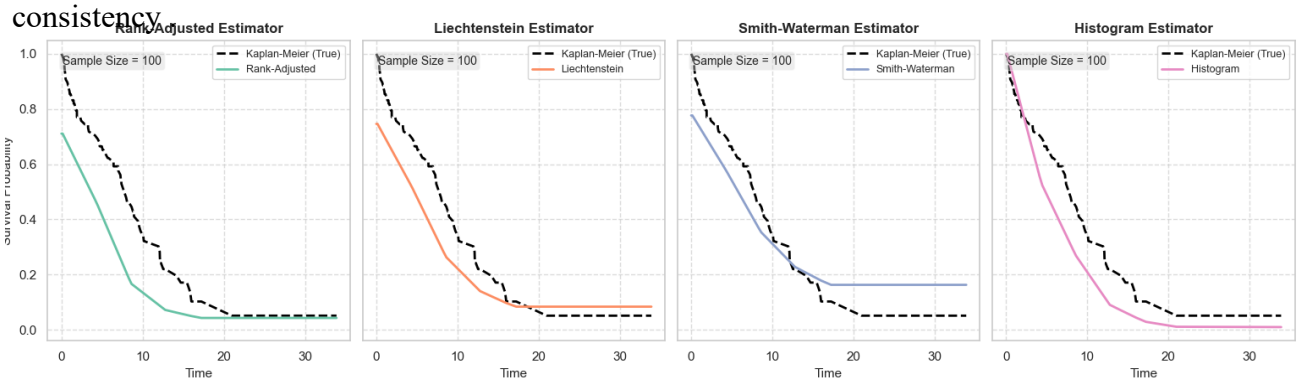

Figure4: Represents the Comparison Curve of Benchmark Model with the Others Estimators at n= 100

Finally, after assuming that the estimate of the survival function represented by the Kaplan-Meier (KM) estimate as a standard model on the basis of which the curve and consistency of those four estimates that were indicated earlier are compared, Figure 3 showed that the rank adjustment method for the graph is excellent by tracing the curve of the real survival function (reference model) the opposite of the estimates of the real survival function at the top of the curve until it decreases at the bottom (at times 20 and above) and this is what I show Also, the curve estimated using Smith-Waterman for the lower part while slightly improved and increased in the upper part to match the average of the real curve, but that the Histogram method of adjusting the rank of the survival function was quite consistent in the upper part at small survival times to remain optimistic and increasing from the real survival estimates while following the same method.

6.2. Discussion

In this part, the results of estimating four methods for adjusting the rank of the reference model are discussed, specifically the Kaplan-Meyer survival function estimation, we rely here in determining

the best method for adjusting the rank after reviewing and evaluating those methods using different performance measures (MSE, AME, RMSE, Bias, Variance, MAPE, and R^2) for different sizes n.

Table 1 Represents the Comparison of Estimates Based on Benchmark model (K-M) at a sample size of 50

Estimator	MSE	MAE	RMSE	Bias	Variance	MAPE	R2
Rank Adjusted Estimator	0.063183005	0.233691	0.251362	-0.23369	0.035849	42.20489	0.153247
Smoothing survival curves	0.037006793	0.171026	0.192371	-0.17103	0.037427	28.68557	0.50405
Smith-Waterman Estimator	0.020245	0.117203	0.142284	-0.1019	0.032611	21.91385	0.728689
Histogram Estimator	0.007745	0.06804	0.088007	-0.06665	0.097301	19.41115	0.896202

Table 1 shows that the estimate of the Rank Adjusted Estimator was weak across all scales, MSE scores from 0.0632, MAE from 0.2337, and RMSE from 0.2514, which indicated high results in estimation errors compared to the rest of the other three estimates, as the bias showed negative by -0.2337, while the moderate variance by 0.0358 reflects the moderate variance, however, the MAPE amounted to 42.20%, and R^2 is 0.1532, these indicate that the first in rank adjustment explains only about 15% of the differences in lifetimes, which puts it in last place in terms of efficiency compared to the rest of the methods.

While it turned out that the Smoothing survival curves estimation method gives better performance than the rank-adjusted estimator, which achieved errors of less MSE than 0.0370, MAE from 0.1710, and RMSE from 0.1924, and that the bias of -0.1710 shows the least systematic underestimation in comparison with the rank-adjusted estimator, while the moderate variance was 0.0374 and MAPE improves to 28.69% and R^2 increases to 0.5041, which means that almost 50% of the variability and prevalence of the lifetime data have been interpreted and calculated, but the Smoothing survival curves estimate is still less efficient than other methods with the exception of the rank-adjusted estimator.

It also turned out that the Smith-Waterman estimate showed a significant improvement in the adjustment of the rank of the survival function estimator, where the performance measures indicated a significant decrease with MSE from 0.0202, MAE from 0.1172, and RMSE from 0.1423, which reflects how low these measures are compared with the rank-adjusted estimator and Smoothing survival curves, while the bias shows a negative that reflected a low systematic error at values of -0.1019, and the variance of 0.0326 indicates the MAPE index is stable while falling to 21.91%, while the R^2 index rises significantly to 0.7287, which indicates the explanation of approximately 73% of the differences in lifetimes, which was put by the second best method for adjusting the rank of a function Stay compared with the estimates above.

The results showed that the Histogram Estimator in the rank adjustment of the survival function estimator has an advantage in performance compared with all the methods used in this study, it gives the lowest MSE of 0.0077, MAE of 0.0680, and RMSE of 0.0880, which indicates these scales indicate the smallest estimation errors and the highest accuracy. While the bias of -0.0667 is the lowest among all the estimates, which leads to the lowest systematic error, but we note that the variance is relatively higher than 0.0973, but on the other hand, it gives high accuracy, as the chart gives the best value by 19.41% and R^2 reaches 90%, which indicated its explanation for differences in life times is captured and explained by the Histogram Estimator, so the Histogram is considered the best in terms of rank adjustment of the estimated Kaplan-Meier survival function at the first sample size n=50.

Table 2 Represents the Comparison of Estimates Based on Benchmark model (K-M) at a sample size of 80

Estimator	MSE	MAE	RMSE	Bias	Variance	MAPE	R2
Rank Adjusted Estimator	0.058992	0.215176	0.242882	-0.21518	0.021755	33.75655	0.112309
Smoothing survival curves	0.034918	0.155507	0.186864	-0.15362	0.024363	22.61983	0.474562
Smith-Waterman Estimator	0.020808	0.117454	0.144251	-0.09196	0.022866	19.69515	0.686882
Histogram Estimator	0.011809	0.097942	0.108668	-0.0963	0.085482	24.12105	0.822304

It turns out from the above table No. 2 that the Histogram Estimator is the most powerful and accurate compared to all other methods, as it achieves the lowest error scales of 0.0118, MAE of

0.0979, and RMSE of 0.1087 which shows that it has the minimum deviation from the truth values of the Benchmark model of survival function, although it records a high variance at 0.0855 with the result corresponding to the lowest amount of bias at (-0.0963), also records the highest R^2 scale at 0.8223, which leads to the interpretation of approximately 83% of the magnitude of the differences in the lifetime data, this means that it can be said here that the Histogram Estimator is the best at the second sample size n=80.

While Smith – Waterman rating in terms of performance was ranked second. Which gave us a strong balance between low error and stable variance, with MSE of 0.0208, MAE of 0.1175, and RMSE of 0.1443. Its bias of -0.0920 is smaller compared to other estimates, while the variance of 0.0229 reflects that the estimation process is consistent to stability. The R^2 value of 0.6869 indicates that 69% of the differences or variation in life times are explained by this estimator, which indicates a significant improvement over the adjusted rank and Smoothing survival curves estimators.

Smoothing survival curves estimators, this estimate shows rather moderate improvements, which put it in third place, but does not exceed in its advantage the Smith-Waterman estimates and the histogram estimate. It turns out that the error scale is — MSE (0.0349), MAE (0.1555), and RMSE (0.1869) — are significantly high, which indicates less accurate estimates. The bias at -0.1536 indicates a systematic reduction, while its divergence (0.0244) is still moderate. The R^2 value of 0.4746 reflects that 47.5% of the variations can be explained which shows how poorly this method modifies the rank of the Benchmark model estimate of the survival function.

The Rank Adjusted Estimator shows us poor performance metrics. The highest recorded are MSE (0.0590), MAE (0.2152), and RMSE (0.2429), which indicates the presence of significant errors in the estimate. Its bias -0.2152 is the largest in magnitude, which indicates the strongest trend of systematic reduction. The variance (0.0218) is the lowest, and the variance has been reduced at the expense of the high amount of bias . With an R^2 value of 0.1123, it explains about 11 percent of the variance in the lifetimes data.

Table 3 Represents the Comparison of Estimates Based on Benchmark model (K-M) at a sample size of 100

Estimator	MSE	MAE	RMSE	Bias	Variance	MAPE	R2
Rank Adjusted Estimator	0.050412	0.216444	0.224526	-0.21644	0.057111	40.08293	0.271116
Smoothing survival curves	0.027861	0.15855	0.166915	-0.15726	0.050628	28.09493	0.597178
Smith-Waterman Estimator	0.014235	0.10766	0.11931	-0.09757	0.04136	21.04068	0.794185
Histogram Estimator	0.014154	0.096193	0.118971	-0.07607	0.108044	24.37327	0.795353

Finally, in Table No. 3 it is shown that the histogram estimate is the best compared with the rest of the estimates and records the lowest MSE (0.01415), MAE (0.09619), and RMSE (0.11897) which indicates high estimation accuracy and minimal deviation from the truth values of the survival function estimate, also indicated the highest R^2 (0.79535), dividing approximately 79.5% of the differences in the data — the strongest explanatory ability among the rest estimates. However, also the variance for this estimate is 0.10804 which resulted in a lower bias which was (-0.07607) and gave the lowest error measures which proved its prestige and efficiency among other methods.

The Smith-Waterman estimator turns out to be the second best performer, as the performance metrics approach the performance of the histogram estimate. It achieves MSE from 0.01424, MAE from 0.10766, and RMSE from 0.11931, which are low and marginally higher than the chart estimate. Its bias (-0.09757) is slightly larger, but it is offset by a significantly lower divergence (0.04136), which indicates more stable fluctuations. Its R^2 value (0.79419) is almost identical to the value of the histogram estimate, explaining approximately 79.4% of the difference in the data. Notably, both the histogram and Smith-Waterman estimators provide robust, nearly convergent estimates, while maintaining a slight advantage at most scales.

It turned out that the Smoothing survival curves estimator ranks third, as it shows moderate improvements in performance measures that are adopted in comparison with the grade-adjusted estimator, but it is less efficient with the above estimators. It registers MSE of 0.02786, MAE of 0.15855, and RMSE of 0.16692, which reflects high estimated errors. The bias (-0.15726) reflects a

significant tendency to systematic underestimation, while the variance (0.05063) is moderate. The R^2 value of 0.59718 indicates that only 59.7% of the differences in the data are explained.
The Rank Adjusted Estimator shows us poor performance metrics. The highest recorded are MSE (0.5041), MAE (0.21644), and RMSE (0.22453), which indicates the presence of significant errors in the estimate. Its bias -0.21644 is the largest in magnitude, which indicates the strongest trend of systematic reduction. The variance (0.05711) is the lowest, and the variance has been reduced at the expense of the high amount of bias. With an R^2 value of 0.27112, it explains about 11 percent of the variance in the lifetimes data.

7.Conclusions

1.It turned out from the results of simulation experiments and for all sample sizes the histogram estimate outperformed all other estimates for the adjustment of the rank of the survival function estimator, which yielded the lowest error measures (MSE, MAE, RMSE) with the least amount of bias and moderate and stable variance, high explanatory power R^2 which nominated it to be the most accurate and reliable.
2.The Smith-Waterman estimator is considered to come in second place in terms of the best performance, as it gives significant improvements in accuracy and explanatory power over the Rank Adjusted and Smoothing survival curves method with a reduction in the size of error and an acceptable and balanced trade-off between bias and variance.
3.The results showed that the Rank Adjusted and the Smoothing survival curves estimator were poor performance measures with high estimation errors, also the magnitude of the bias is large and the lowest values of R^2 in different sample sizes which made them the least effective estimates in this comparison.
4.The increase in sample sizes leads to an improvement in the performance of estimates, which led to a reduction in error measures, and this is fully consistent with the statistical theory (MSE, MAE, RMSE), as well as the obvious increase in R^2 when the sample size moves from 50 to 100.
5.The histogram method maintained its overall superiority compared to the rest of the other estimates, despite its relatively high variance, which came in line with the trade-off that favors low bias over variance, which leads to better prediction accuracy.
6.We note that the performance advantage gap between the histogram estimate and the Smit-Waterman estimates begins to narrow at large sample sizes (N=100), where increasing the sample size leads to the achievement of both estimates to similar accuracy and explanatory power. This indicates that by increasing the sample size or at large sample sizes, the estimators are strong competitors.

8.Recommendations

1.The priority of the histogram estimate is the applied aspect that requires high accuracy in estimating the survival function, due to its superior performance in reducing error performance measures errors (MSE, MAE, RMSE) and maximizing the explanatory power of differences in the lifetime data generated in the simulation aspect, so here we recommend using histogram estimate in case there is a need for more accurate estimates with a trade-off of less bias and relatively greater variance.
2.We take into account the Smith-Waterman estimator as it gives balanced performance measures with different sample sizes, as Smith-Waterman provides a good balance between stability and accuracy, and this is indicated by the trade-off between bias and variance, especially at medium and large sample sizes in applied medical.
3.The selection of large sample sizes in clinical trial applications for the purpose of achieving high accuracy and efficiency, where it is recommended to use sample sizes of 100 and larger as explained in the analysis where the histogram estimates match with the Smith-Waterman.

9. References

1. Allen, J. D., Savadatti, S., & Gurmankin Levy, A. (2009). The transition from breast cancer 'patient'to 'survivor'. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*, *18*(1), 71-78. https://doi.org/10.1002/pon.1380.

2. Chan, S., & Airoldi, E. (2014). A consistent histogram estimator for exchangeable graph models. *Proceedings of the 31st International Conference on Machine Learning*, 32, 208–216. PMLR. https://proceedings.mlr.press/v32/chan14.htmlProceedings of Machine Learning Research+1Proceedings of Machine Learning Research+1
3. Chen, E. J., & Kelton, W. D. (2001). Quantile and histogram estimation. *Proceedings of the 2001 Winter Simulation Conference*, 1, 451–459. IEEE. https://informs-sim.org/wsc01papers/059.PDFINFORMS SIM
4. Chen, L., Feris, R., & Turk, M. (2008). Efficient partial shape matching using Smith-Waterman algorithm. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1–6. IEEE. https://rogerioferis.com/publications/FerisNordia08.pdfRogerio Feris
5. Collett, D. (2023). *Modelling survival data in medical research* (4th ed.). Chapman and Hall/CRC. https://www.routledge.com/Modelling-Survival-Data-in-Medical-Research/Collett/p/book/9781032252858Routledge
6. Comet, J. P., Aude, J. C., Glémet, E., Risler, J. L., Hénaut, A., Slonimski, P. P., & Codani, J. J. (1999). Significance of Z-value statistics of Smith–Waterman scores for protein alignments. *Computers & Chemistry*, 23(3–4), 317–331. https://www.sciencedirect.com/science/article/pii/S009784859900008XScienceDirect
7. Etikan, I., Abubakar, S., & Alkassim, R. (2017). The Kaplan-Meier estimate in survival analysis. *Biometrics & Biostatistics International Journal*, 5(2), 55–59. https://medcraveonline.com/BBIJ/the-kaplan-meier-estimate-in-survival-analysis.htmlSCIRP+2MedCrave Online+2MedCrave Online+2
8. Efron, B. (1967). The Jackknife, the Bootstrap and Other Resampling Plans. *SIAM (Society for Industrial and Applied Mathematics)*. DOI: https://doi.org/10.1137/1.9781611970319
9. Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L^2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4), 453–476. https://link.springer.com/article/10.1007/BF01025868SpringerLink
10. Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research*, 1(4), 274. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/
11. Härdle, W. (2012). *Smoothing techniques: With implementation in S*. Springer Science & Business Media. https://link.springer.com/book/10.1007/978-1-4612-4432-5SpringerLink
12. Jung, S. H. (2008). Sample size calculation for the weighted rank statistics with paired survival data. *Statistics in Medicine*, 27(17), 3350–3365. https://pubmed.ncbi.nlm.nih.gov/18205148/PubMed
13. Khajeh-Saeed, A., Poole, S., & Perot, J. B. (2010). Acceleration of the Smith–Waterman algorithm using single and multiple graphics processors. *Journal of Computational Physics*, 229(11), 4247–4258. https://www.sciencedirect.com/science/article/pii/S0021999110000823
14. Kontkanen, P., & Myllymäki, P. (2007). MDL histogram density estimation. *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, 2, 219–226. PMLR. https://proceedings.mlr.press/v2/kontkanen07a.htmlProceedings of Machine Learning Research+1Proceedings of Machine Learning Research+1
15. Kunitomo, N., & Matsushita, Y. (2008). Improving the rank-adjusted Anderson-Rubin test with many instruments and persistent heteroscedasticity (No. CIRJE-F-588). CIRJE. https://ideas.repec.org/p/tky/fseres/2008cf588.htmlIDEAS/RePEc
16. Kim, C., Park, B. U., Kim, W., & Lim, C. (2003). Bezier curve smoothing of the Kaplan-Meier estimator. *Annals of the Institute of Statistical Mathematics*, *55*(2), 359-367.
17. Lehto, A., Cherikh, L., Susi, A., Shvartsman, K., Peterson, L., Nylund, C. M., & Brown, J. (2025). Female permanent contraception in the Military Health System after the Dobbs v. Jackson Women's Health Organization decision. *O&G Open*, 2(3), e079. https://journals.lww.com/ogopen/fulltext/2025/06000/female_permanent_contraception_in_the_military.2.aspxLippincott Journals
18. Leys, C., & Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology*, 46(4), 684–688. https://www.sciencedirect.com/science/article/pii/S002210311000034XScienceDirect
19. Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports,* *50*(3), 163–170. https://pubmed.ncbi.nlm.nih.gov/5910392/

20. Marron, J. S., & Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *Journal of Multivariate Analysis, 20*(1), 91–113. https://doi.org/10.1016/0047-259X(86)90021-7
21. Morris, T. P., Jarvis, C. I., Cragg, W., Phillips, P. P., Choodari-Oskooei, B., & Sydes, M. R. (2019). Proposals on Kaplan–Meier plots in medical research and a survey of stakeholder views: KMunicate. *BMJ Open, 9*(9), e030215. https://doi.org/10.1136/bmjopen-2019-030215
22. Mott, R. (1992). Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bulletin of Mathematical Biology, 54*(1), 59–75. https://doi.org/10.1016/S0092-8240(05)80176-4
23. Nematolahi, S., Nazari, S., Shayan, Z., Ayatollahi, S. M. T., & Amanati, A. (2020). Improved Kaplan-Meier estimator in survival analysis based on partially rank-ordered set samples. *Computational and Mathematical Methods in Medicine, 2020*, Article 7827434. https://doi.org/10.1155/2020/7827434
24. Nieto, F. J., & Coresh, J. (1996). Adjusting survival curves for confounders: A review and a new method. *American Journal of Epidemiology, 143*(10), 1059–1068. https://pubmed.ncbi.nlm.nih.gov/8629613/
25. Wu, Y., & Kolassa, J. (2024). Interval-specific censoring set adjusted Kaplan–Meier estimator. *Journal of Applied Statistics*, *51*(12), 2436-2456. DOI: 10.1111/j.0006-341x.2002.00439.x