An Analysis of the Kaplan-Meier Estimator and the Consequences of Censoring

Holly Hammons, Jessica Kennedy, Roy Lundeen, Richard Sternesky

***Abstract*—In 1958, Kaplan and Meier developed a nonparametric statistic for estimating the survival function from lifetime data. The Kaplan-Meier estimator is known for being capable of accounting for censored data, which allows it to be an accurate estimator for use in clinical trials. To better understand the Kaplan-Meier estimator, our team explores its statistical properties in detail and its numerous applications. In addition, we analyze how different models of censoring affect the estimated survival function from the Kaplan-Meier estimator and how to best mitigate these effects. Furthermore, we compare the effects of censoring when using the Kaplan-Meier estimator to another well-known method, the Nelson-Aalen estimator. Our team found that the characteristic randomness of natural survival data produces jagged patterns in true survival curves, which can more accurately be estimated by the Kaplan-Meier estimator in the presence of uniform censoring than in the presence of random censoring. Moreover, the censored survival curve generated from the Nelson-Aalen estimator is found to follow the uncensored survival curve more closely than the Kaplan-Meier estimator, however, the differences between them are small enough to be considered trivial. Therefore, the Kaplan-Meier estimator is a suitable tool for use in estimating survival probabilities in clinical trials where the presence of censored data can exist.**

***Keywords*—** Kaplan-Meier estimator, Survival function, Right-censored

I. Introduction

SURVIVAL time can be defined as the time starting from a defined point to the occurrence of a given event, while survival analysis is defined by the analysis of such data in groups [10]. Survival analysis is used a lot in environments such as clinical trials in which survival time of a patient of a particular disease can be estimated.

However, in these settings, survival analysis can quickly become complicated if patients are no longer a part of a trial due to reasons other than recovery or death, such as patients prematurely leaving or becoming uncooperative [10]. Circumstances such as these can be labeled as right-censored observations, and must be considered when estimating survival time.

In addition to those subjects that prematurely leave a trial, there must be consideration about subjects that begin the trial later, after the time of the first event. While this data will be inconsistent with previous data, it must still be considered as a part of the entire survival analysis [10].

While there are many ways of studying survival analysis, this paper will focus on the Kaplan-Meier estimator. The Kaplan-Meier estimator is a nonparametric statistic that is used to estimate the survival function [3].

This method for estimating a survival function is accurate in the presence of censored data. At every time interval of an event occurring, the survival probability is calculated by subtracting from one the division of the number of events that occur by the number of events that have yet to happen. If a patient drops out of a study, their data becomes right-censored and is no longer considered in the denominator. [10]. Using the multiplication rule, each of the probabilities, from every time interval, will be multiplied together to estimate the entire survival function [10].

II. Theory and Derivation

A. Background Information

In order to understand the Kaplan-Meier estimator, one must be familiar with the survival function. Let a lifetime t be a continuous random variable with probability density function and cumulative distribution function shown below [15],

$$\left.\begin{array}{ll} pdf: & f(t) \\ cdf: & F(t)=P(T\leq t) \end{array}\right\} \quad F(0)=P(T=0)$$

Therefore, the survival function can be defined as [15],

$$S(t)=1-F(t)=P(T>t) \quad for \quad t>0. \quad (1)$$

The Survival function describes the probability that a system, individual, or component will not fail after a given time [4].

B. The Kaplan-Meier Estimator

There are three assumptions that must be made for the Kaplan-Meier estimator [10]. First, we must assume that patients who are reported as censored have the same survival probabilities as those still in the trial. Second, we must assume that survival prospects remain the same for those who join a trial at the beginning and after time has passed. Lastly, an event is assumed to have occurred at each time that it is specified. There are four possibilities for each interval that must be considered [7]:

1) There is no death or censored data to report. In this case, the survival probability is estimated to be approximately one.
2) There is at least one censored data reported but no death reported. Using the assumptions above, we can assume the conditional probability of surviving this interval to be approximately one.
3) There is at least one death reported but there is no censored data reported. In this case, the estimated conditional probability must be approximately,
$$1-\frac{d}{n}, \quad (2)$$
where d is the number of deaths at that specific time, and n is the number of those who have not yet had a death or have not been censored.
4) There are both at least one death reported and no censoring data reported. Again, using the assumptions made above, the estimated conditional probability can be written as equation (2).

Therefore, using the multiplication rule, the Kaplan-Meier estimator can be used to estimate a survival function as [3],

$$\hat{S}(t)=\prod_{i:t_i\leq t}(1-\frac{d_i}{n_i}), \quad (3)$$

where t_i is the time that at least one event happens, d_i is the number of deaths at that specific time, and n_i is the number of those who have not yet had a death or been censored. It is worth being clear that, based on the assumptions made above, censored data is no longer considered a part of the denominator, n_i, at the time they leave a study [10].

C. Estimated Sample Mean

Kaplan and Meier [3] note that the mean of a nonnegative random variable is equal to the area under the corresponding survivorship function and is written as,

$$\hat{\mu} = \int\limits_{0}^{\infty} \hat{S}(t)\, dt. \quad (4)$$

However, if $S(t)$ is not defined everywhere, this will cause the mean to be undefined [3]. With an assumption that the probability of an indeterminate result is small, $S(t)$ is practically unbiased [3]. Therefore, the estimated mean of the survival function can simply be calculated by finding the weighted average of all of the values that T can take, where the weights are given by the probabilities that T takes each of those values [11]. In the case of the survival function, T will be an interval from the time of one event occurring to the next. Therefore, in the instance that the probability of an indeterminate result is small, the estimated mean can be found by [3],

$$\hat{\mu} = \sum_{i:t_j<t} \hat{S}(t_i)(t_{i+1} - t_i). \quad (5)$$

In the instance that the probability of an indeterminate result is high, instead of estimating the mean itself, the "mean life limited to a time L" can be estimated [3]. If this method of estimating the mean is chosen, the estimate of $S(L)$ should be noted, along with the estimated mean ending at time period L [3]. Instead of the simple weighted average of the entire span of intervals, the average will only be found up until time L. The final piece is adding this average to the product of the interval from the last time period to time L, with the probability of surviving that interval. Therefore, when probability of an indeterminate result is high, the estimated mean can be found by [3],

$$\hat{\mu} = \sum_{i:t_j<t} \hat{S}(t_i)(t_{i+1} - t_i) + \hat{S}(t_L)(L - t_{i+1}). \quad (6)$$

D. Estimated Variance

As the Kaplan-Meier estimator estimates a survival function, the variance must be an estimate as well. While many researchers have developed different ways of estimating the variance of the Kaplan-Meier estimator, the most common method is by using the Greenwood variance estimator [2].

In order to derive the Greenwood variance estimator, the Delta method must be explained. By using of the Central Limit Theorem [1], we can approximate that the sum of a large number of independent random variables essentially has a normal distribution. Consider that Y is a random variable with mean μ and variance σ^2,

$$Y_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right). \quad (7)$$

The Delta method states that, if g is defined as a smooth function with no discontinuous jumps in values, and g is it's first derivative, then $g(Y_n)$ can be approximated as [7],

$$g(Y_n) \approx \mathcal{N}\left(g(\mu), \frac{g'(\mu)^2\sigma^2}{n}\right). \quad (8)$$

Note that equation (8) is only an approximation because, when looking at it's Taylor Series expansion, it becomes an infinite sum [1]. One can derive the first order Taylor Series expansion of this result, which allows the delta method to estimate the variance of a function of a random variable $f(X)$ by [9],

$$\begin{aligned} Var(f(X)) &\approx Var[f(\mu) + f'(\mu)(X - \mu)] \\ &\approx (f'(\mu))^2 Var(X). \end{aligned}$$

Now, we'll consider a function of $\hat{S}(t)$ as the log of the survival function,

$$log(\hat{S}(t)) = \sum_{j:\tau_j<t} log(1 - \frac{d_i}{n_i}). \quad (9)$$

Let the probability of surviving a given interval from $[t_i, t_{i+1})$ be defined as $\hat{p}_{t_i}$. By assuming that the probability of surviving t_{i+1} is only dependent on surviving to time t_i, we can define the variance as [9],

$$Var(log(\hat{S}(t)) = \sum_{i:t_i\leq t} Var(log(\hat{p}_{t_i})). \quad (10)$$

Using the delta method, equation (10) can be rewritten as [9],

$$Var(log(\hat{p}_{t_i})) \approx \left[\frac{1}{\hat{p}_{t_i}}\right]^2 Var(\hat{p}_{t_i}). \quad (11)$$

To find $Var(\hat{p}_{t_i})$, we must assume that the number of individuals remaining in a risk set, accounting for both deaths and censored events, denoted as, $(n_{t_i} - d_{t_i})$, to be a binomially distributed random variable [9]. This random variable is described by its number of trials, denoted as n_{t_i}, and the probability of success, denoted as $\hat{p}_{t_i}$. By assuming a binomial distribution, its expectation and variance can be defined as [9],

$$\begin{aligned} E(n_{t_i} - d_{t_i}) &= n_{t_i}\hat{p}_{t_i} \\ Var(n_{t_i} - d_{t_i}) &= n_{t_i}\hat{p}_{t_i}(1 - \hat{p}_{t_i}) \end{aligned}$$

Now, the $Var(\hat{p}_{t_i})$ can be found by [9],

$$\begin{aligned} Var(\hat{p}_{t_i}) &= Var\left(1 - \frac{d_{t_i}}{n_{t_i}}\right) \\ &= Var\left(\frac{n_{t_i} - d_{t_i}}{n_{t_i}}\right) \\ &= \left(\frac{1}{n_{t_i}}\right)^2 Var(n_{t_i} - d_{t_i}) \\ &= \left(\frac{1}{n_{t_i}}\right)^2 n_{t_i}\hat{p}_{t_i}(1 - \hat{p}_{t_i}) \\ Var(\hat{p}_{t_i}) &= \frac{\hat{p}_{t_i}(1 - \hat{p}_{t_i})}{n_{t_i}}. \quad (12) \end{aligned}$$

Thus, the variance from equation (12) can be plugged into equation (11) to find $Var(log(\hat{p}_{t_i}))$ from the delta method by [9],

$$\begin{aligned} Var(log(\hat{p}_{t_i})) &\approx \left[\frac{1}{\hat{p}_{t_i}}\right]^2 \left(\frac{\hat{p}_{t_i}(1 - \hat{p}_{t_i})}{n_{t_i}}\right) \\ &\approx \frac{1 - \hat{p}_{t_i}}{\hat{p}_{t_i} n_{t_i}} \\ &\approx \frac{d_{t_i}}{n_{t_i}(n_{t_i} - d_{t_i})}. \quad (13) \end{aligned}$$

Therefore,

$$Var(log(\hat{S}(t)) = \sum_{i:t_i\leq t} \frac{d_{t_i}}{n_{t_i}(n_{t_i} - d_{t_i})}. \quad (14)$$

Equation (14) can now be used to find $\hat{Var}\left(\hat{S}_{t_i}(t)\right)$. Let $X = log(\hat{S}(t)$, and by further use of the delta method [6],

$$\begin{aligned} \hat{Var}\left(\hat{S}_{t_i}(t)\right) &= \hat{Var}(exp(X)) \\ &= [exp(x)]^2 Var(X) \\ &= [\hat{S}_{t_i}(t)]^2 \sum_{i:t_i\leq t} \frac{d_{t_i}}{n_{t_i}(n_{t_i} - d_{t_i})}. \end{aligned}$$

Therefore, the variance of $\hat{S}(t)$ can be defined as [9],

$$\hat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i\leq t} \frac{d_i}{n_i(n_i - d_i)}. \quad (15)$$

III. Effect of Censoring Distributions on the Kaplan Meier Estimator

As a prerequisite to our analysis on censorship, we required a robust model for survivorship which could emulate real world data without losing the inherent randomness of novel survivorship applications. One of the main difficulties in generating real world survival data is the necessity to ground randomly or pseudo-randomly generated raw numbers in a pre-determined hazard function, most often the exponential, Weibull, or Gompertz models [8]. This assumption of foreknowledge undermines the characteristic irregularities that real-world survivorship curves so often exhibit, and can skew the results of analysis techniques or observations such as those conducted in this paper. To mitigate this common pitfall, Harden and Kropko's [8] 'coxed' R library was used to generate a cubically smoothed failure CDF, which we then used to derive the failure PDF curve as shown in Figure 1. From the failure density function, survivorship data can be created by generating duration data according to the relative probability of survival at any given point prior to the CDF reaching approximately 1. This duration data is then used to model the survivor and hazard functions by calculating the frequency of 'deceased' subjects at every time point prior to a failure rate of 100%, which is equal to the number of durations less than or equal to the current time point, and extrapolating between unrepresented time points [8].

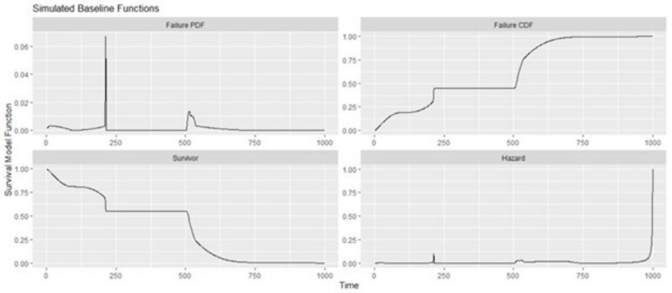


Fig. 1. The results of simulating survivor data using the algorithm set out by Harden and Kropko [8]

Visually, the survival curve of the generated data set used in our initial simulations portrays the 'jagged' and almost discontinuous behavior of real-world survival curves at several points in time. The spikes present in the failure probability density function also display the random nature of this data set, which likely could not be observed in data generated via a pre-determined hazard function.

Using the dataset and survivor curve developed in R, we can apply a variety of censoring distributions to the Kaplan-Meier estimator. For the purposes of our simulation, we emulated censoring by sampling the data at certain time points based on pre-determined distributions. Effectively, we suppose that the data generated in R is a true and accurate representation of the particular survival curve we are attempting to model. Then, we sample this data with a given sampling ratio (the ratio between sampled data points and original data points, which we arbitrarily chose as 1/10) and distribution. This is a realistic model of the unintentional source of error which results from not having access to the full sample space of a given survival application. In disease trials or machine testing, researchers can never collect survival data on everyone with the disease they are studying, nor can engineers access the lifetime data of every product they manufacture. By comparing a number of these simple distributions, we are able to observe how different levels of censoring can obscure the true probability of survival.

We used the raw data to first simulate the Kaplan-Meier estimator using the MATLAB ecdf function, which produces a Kaplan-Meier estimation using the given survival data. We then created uniform, random normal, and random uniform censored versions of the original dataset, and plotted them against the survivorship function and uncensored estimator as shown in Figure 2. We can broadly analyze the impact of these various distributions on the accuracy of the Kaplan-Meier estimator using visual inspection, which reveals the low impact of uniform censoring on the estimator as compared to normal random and uniform random censoring. To further substantiate this conclusion, we use the built-in Matlab functions for determining the expectation and standard deviation of each estimate, comparing these values to the original data's characteristics to determine their correspondence. Summarized in Table 1, the expectation and standard deviation of the uniformly distributed censoring scheme closely match those of the uncensored data as expected, with the expectations differing by only 5 time units and the standard deviations by 2.

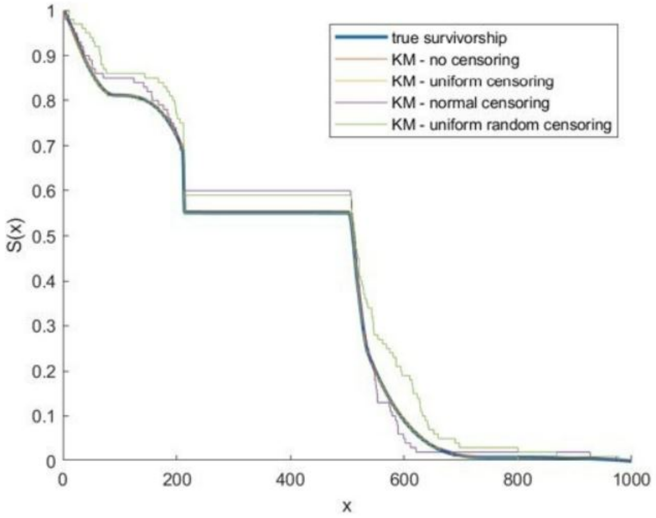
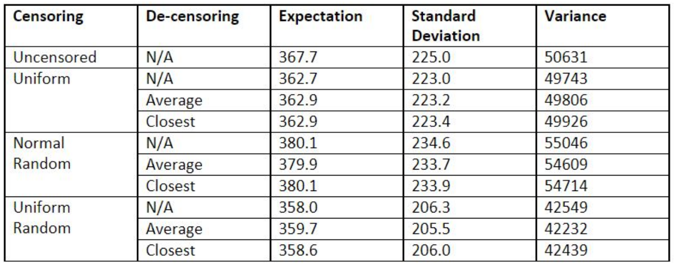


Fig. 2. By plotting the distributions together, a more comprehensive view of the differences in each curve is achieved. The application of purely uniform censoring has little effect on the Kaplan-Meier estimator.

Compared to the uncensored and uniformly censored estimators, the Kaplan-Meier curve with normal censoring takes on a rougher form, with steeper drops at larger intervals. This is the result of a decrease in the number of eligible patients over particular intervals, but an unchanging number of deaths at any given time t. We see similar effects when random uniform censoring is added, as well as a dramatic difference from the uncensored curve.



Censoring	De-censoring	Expectation	Standard Deviation	Variance
Uncensored	N/A	367.7	225.0	50631
Uniform	N/A	362.7	223.0	49743
	Average	362.9	223.2	49806
	Closest	362.9	223.4	49926
Normal Random	N/A	380.1	234.6	55046
	Average	379.9	233.7	54609
	Closest	380.1	233.9	54714
Uniform Random	N/A	358.0	206.3	42549
	Average	359.7	205.5	42232
	Closest	358.6	206.0	42439

Fig. 3. Once taken into Matlab, we can calculate the expectation, standard deviation, and variance of each survival curve using built-in Matlab functions.

As a means of mitigating the effects of our censoring on the Kaplan-Meier estimator, we next sought to de-censor the sampled data by applying two different algorithms to our estimates. In the first algorithm, we rectify the sampled data's tendency to disrupt stable intervals of the estimator by randomly choosing outlier events. We accomplish this task by simply averaging the 100 data points surrounding the selected value. This algorithm has the effect of 'smoothing' the Kaplan-Meier estimator, and routinely decreased the difference between real and censored values of the expectation and variance for all distributions save the uniform distribution. Next, in an attempt to stay more truthful to the original data, we applied a 'closest fit' algorithm, which takes the value closest to the average of the 100 surrounding data points within the set of these 100 points. This would theoretically reduce the effect of outliers even more dramatically, as the influence of outliers would be eliminated when the only 'average' real data points are selected from. This method proved to be less effective in mitigating the impact of censoring on the generated data as seen in Figure 3, possibly due to the jagged nature of the simulated survivorship curve to begin with. In a system with near discontinuous behavior (or cusp creating behavior on a derivative plot) at points, defining time intervals in the shape of the survivorship curve can be dramatically altered with an algorithm that steeply disfavors outlier behavior, with this algorithm creating a 'lag' of sorts in the Kaplan-Meier estimate.

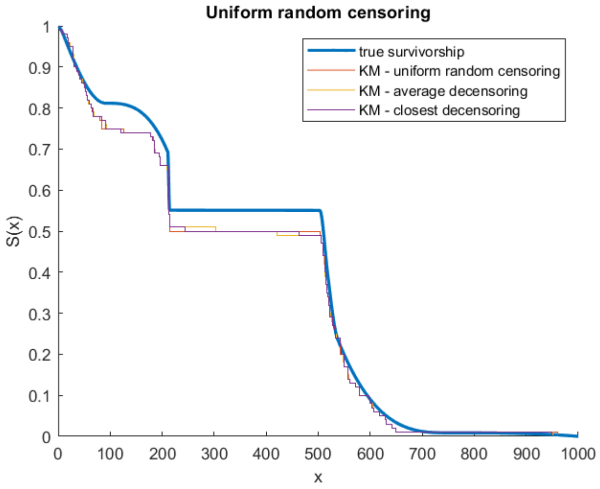


Fig. 4. The impact of censoring and de-censoring for the uniform random distribution can be seen in this plot of all Kaplan-Meier estimates compared to the true survival curve.

IV. Comparison: The Nelson-Aalen Estimator and Censoring

The Nelson-Aalen estimator is known for being a popular non-parametric estimate for the cumulative hazard rate function. The cumulative hazard rate function, $H(t)$, can be found by integrating the hazard function over a period of time, and is related to the survival function by [6],

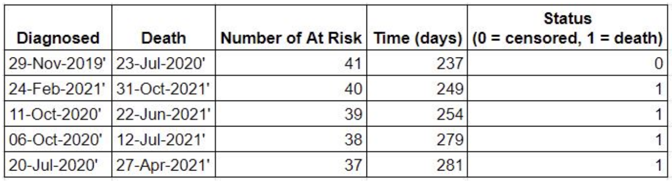
$$S(t) = e^{-H(t)} \quad (16)$$

The Nelson-Aalen estimator is a non-parametric statistic that is used for estimating the cumulative hazard rate function and can be defined as [13],

$$\hat{H}(t) = \sum_{i:y_i \leq t} \frac{d_i}{n_i} \quad (17)$$

Given that both the survival function and the hazard function can be estimated using the same data sets, and that they can be related to each other using equation (16), they can both be used as a comparison point between how censoring effects each of their estimated survival curves.

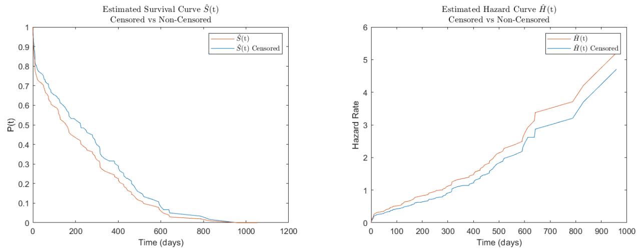
In order to develop a general method for comparing the effect of censoring on these two estimators, we created our own arbitrary data set using MATLAB. Using a random uniform distribution, the diagnosis dates and death dates were generated for each hypothetical patient. Then, the time between these dates was calculated, allowing the patients to be sorted by their times of survival. We then produced an additional data set with the same dates of diagnosis and death, however, each patient was then labeled to be either censored non-censored by the time that they left the clinical trial. The censoring was decided by use of a normally distributed random number with a mean of 0.6 and a standard deviation value of 0.1. The table below shows 5 entries of the randomly generated data, with the addition of a column denoting the status of censorship.



Diagnosed	Death	Number of At Risk	Time (days)	Status (0 = censored, 1 = death)
29-Nov-2019'	23-Jul-2020'	41	237	0
24-Feb-2021'	31-Oct-2021'	40	249	1
11-Oct-2020'	22-Jun-2021'	39	254	1
06-Oct-2020'	12-Jul-2021'	38	279	1
20-Jul-2020'	27-Apr-2021'	37	281	1

Fig. 5. Randomly Generated Arbitrary Data Set

Using this data set, we first calculated the estimate of the survival function for both the censored and not censored data by using the Kaplan-Meier estimator. Then, we calculated the estimate of the hazard function for both the censored and not censored data by using the Nelson-Aalen estimator.



(a) (b)

Fig. 6. (a) Estimated Survival Curve from Kaplan-Meier Estimator
(b) Estimated Hazard Curve from Nelson-Aalen Estimator

Using equation (16), we related the hazard function to the survival function. It is obvious that, while there are minor discrepancies between the two survival curves as time increases, they generally follow the same trend, for both censored and not censored data. However, there are larger discrepancies between the survival curves with respect to if the data was censored or not.

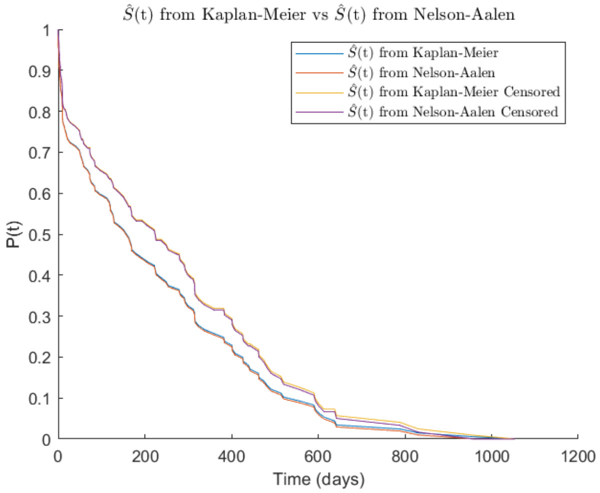


Fig. 7. Estimated Survival Curves from Kaplan-Meier and Nelson-Aalen Estimators, Censored vs Non-Censored

Below is a graph representing the percent difference between the Kaplan-Meier and Nelson-Aalen estimated survival curves of censored and non-censored data. As the number of events increase, it is natural for both of the survival curves with censoring to deviate further from the survival curves without censoring. Most notably, we noticed a dramatic increase as the curves neared their final events. We propose that this is due to the dramatic decreases in each survival curve towards the end of a trial, as the probability of survival naturally falls towards zero. While the differences between the two are small, it is shown that the Kaplan-Meier estimator has a slightly larger difference between the censored survival curve and the non-censored survival curve as the number of events increases.

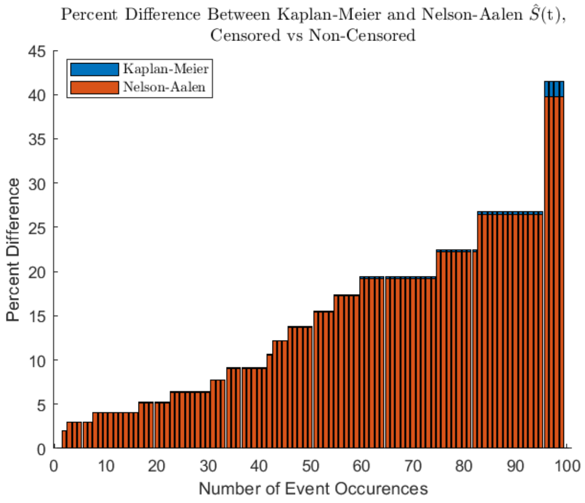


Fig. 8. Percent Difference Between Censored vs Non-Censored Estimated Survival Curves from Kaplan-Meier and Nelson-Aalen Estimators

V. Applications and Extensions

As mentioned prior, the predominant application of the Kaplan-Meier estimator is in clinical trials to analyze the differences between a control and a study group. For example, a study of the Health Insurance Plan of Greater New York in 1977 researched the efficacy of breast cancer screening in women aged 40 to 69. Researchers used the Kaplan-Meier method to determine the survival function for both the control and the study groups, each with approximately 31,000 women participants [14]. The control group continued their routine medical care, and the study group received an annual breast examination. After the first year of the study, there were 30% more deaths in the control group than the study group, as shown in the table below [14]. The corresponding survival probability graph, derived from the Kaplan-Meier estimator, is shown below [20].

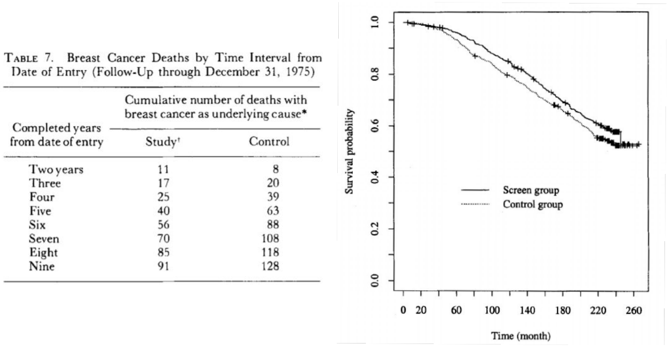


Table 7. Breast Cancer Deaths by Time Interval from Date of Entry (Follow-Up through December 31, 1975)

Completed years from date of entry	Cumulative number of deaths with breast cancer as underlying cause*	
	Study†	Control
Two years	11	8
Three	17	20
Four	25	39
Five	40	63
Six	56	88
Seven	70	108
Eight	85	118
Nine	91	128

(a) (b)

Fig. 9. (a) Breast Cancer Trial Results [14] (b) Survival Probability Graph [20]

As an extension to this, an additional study from 2001 used the original study's data to calculate a range of test statistics. By doing so, researchers can better interpret the differences between the control and study groups when their results are relatively similar. Researchers noticed that the first 40 months of follow-up yielded a similar survival probability between the control and the study groups. The base of this test statistic is given by,

$$K_f(t) = f\hat{S}(t-)\frac{\hat{C}_1(t-)\hat{C}_2(t-)}{p_1\hat{C}_1(t-) + p_2\hat{C}_2(t-)} \quad (18)$$

where $\hat{C}_j(t)$ is the probability that a given time is greater than the censored time, and p_j represents the number of fatalities in the group over the number of participants (not including censored participants). The subscripts 1 and 2 distinguish the study and the control groups. The $\hat{S}(t-)$ represents the survival function up to the specific time. The function f was selected as a log-rank function, taking the form,

$$f(t) = t^{\rho-1}(1-t)^{\gamma-1} \quad (19).$$

Rho and gamma were tested in different combinations on the range [1,3]. $f(t)$ was then calculated at different survival probabilities for the control and study groups, and the covariance was determined. This is plotted in the graph below [20]. This graph shows how, when $\gamma = 1$, the test statistics follow a decreasing exponential trend. All other combinations follow a curve very similar to the Erlang distribution. The implementation of a range of test statistics allows researchers to better interpret their data when the difference between the control and study groups is not obvious [20].

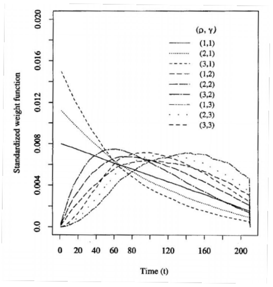
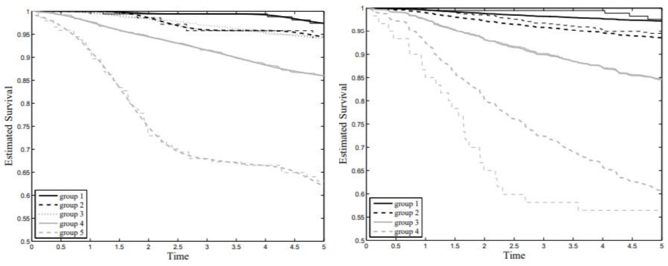


Fig. 10. Implementation of Test Statistics [20]

The Kaplan-Meier estimator is commonly used as a base component for extensive medical data modeling. A study through the Electrical Engineering department at KU Leuven performed an analysis on breast cancer patients by using the Kaplan-Meier estimator as a foundation. Their approach introduced machine learning principles to cluster participants based on a score. Their computational model then divided the study participants into groups, where the number and score ranges are determined by the clustering algorithm [17].

Each cluster is treated as its own dataset and is analyzed using the Kaplan-Meier estimator. KU Lauven derived an "inverse probability of censoring" function as$\hat{F}_{RR} = 1 - \hat{S}_{KM}$, where S_{KM} is the survival function derived from the Kaplan-Meier estimator. Additionally, they applied their own smoothing algorithm in order to generate a continuous graph from the discrete result. The output of their smooth graph is the function $\hat{G}$. Figure 11 shows their results when applying their data to a pool of breast cancer trial participants. Each participant was assigned a risk index, which was based on the number of lymph nodes, tumor grade, and additional hormone indicators [17].

These graphs show the effectiveness of their clustering model on grouping participants together based on risk because of the distinct difference in the survival functions across the groups. Though, there could be room for refinement in their algorithm because the difference between groups 2 and 3 is very small and follow similar trends. Further exploration of this de-censoring algorithm was conducted by Willems et. al. in their analysis of dependent censoring. In this work, researchers applied an algorithm in which non-censored data is



(a) (b)

Fig. 11. Effects on Clustering [17] (a) Group A Survival Curve (b) Group A Survival Curve

weighted more heavily than censored data using an inverse proportion scheme [19]. This algorithm more usefully mitigates the impacts of censoring by relying only on given data, which is a significant advantage when ‘excess’ data cannot simply be thrown out as it was in our de-censoring models.

One final method of reducing the negative impact of censoring on data was introduced in Falcaro and Capenter’s study on correcting bias due to missing data [8]. In this solution, plausible data sets are created from existing data and then combined with the original set at various levels. This method for de-censoring allows for uncertainty in data sets, and strongly relates to the type of censoring we introduced in our own approaches. A practical approach to correcting for uncertainty caused by a lack of raw data, multiple imputation offers a reliable method for further bolstering confidence in recorded survival curves without undermining original results [8].

VI. Conclusion

The Kaplan-Meier estimator is a useful tool for estimating the survival function from lifetime data in the presence of censoring. While its survival probabilities are only estimates, researchers can find usefulness in its results and statistical information such as its mean and variance.

Before analyzing the effects of censorship on the Kaplan-Meier estimator, we found it necessary to use random test data that closely resembled real-world data. This provides a level of confidence in the integrity of our data, and prevents bias associated with using common distributions. Based on our tests, we conclude that the Kaplan-Meier estimator is resilient to most forms of censoring. We observed mild deformation in the survival curves given normal or random uniform censoring, as well as a “jaggedness,” characteristic of the use of time intervals in our code. It was evident in our data that uniform censoring did not influence the estimator, and the percentage difference in expectation from the uncensored estimator remained below approximately 2-3%. We conclude that the Kaplan-Meier estimator is proficient at mitigating censoring, and the nature of its equation inherently reduces its effects. This result is supported by the proximity of the expectation values from each curve, as the drastic visual differences are not reflected in the expectation values. This indicates that an estimator with a common censoring distribution is nearly as effective as an uncensored Kaplan-Meier estimator, with an expectation within 4% of the original.

By comparing the Kaplan-Meier and Nelson-Aalen estimators, we could determine how each survival curve behaves in the presence of censored and non-censored data. We found that, as the number of events increase, the Kaplan-Meier estimator shows a slightly larger difference between the censored survival curve and the non-censored survival curve than the Nelson-Aalen estimator. However, given that the percent difference between each estimator rarely exceeds 10% until roughly the final 10 events, we can conclude both estimators to be nearly as effective in modeling a survival curve from random test data.

Appendix A

MATLAB code, including equations used and defined variables, can be found as an attached copy to this report.

References

[1] A. Papanicolaou, “Taylor Approximation and the Delta Method,” Academia, Apr. 2009.

[2] C. B. Borkowf, “A simple hybrid variance estimator for the Kaplan–Meier survival function,” STATISTICS IN MEDICINE, Nov. 2004.

[3] E. L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” Journal of the American Statistical Association, vol. 53, no. 282, pp. 457–481, 1958.

[4] J. Abdullahi, U. Kabir Abdullahi, T. Godfrey Ieren, D. Adugh Kuhe, and A. Abubakar Umar, “On the Properties and Applications of Transmuted Odd Generalized Exponential-Exponential Distribution”, AJPAS, vol. 1, no. 4, pp. 1-14, Oct. 2018.

[5] J. J. Harden and J. Kropko, “Simulating Duration Data for the Cox Model,” Political Science Research and Methods, vol. 7, no. 4, pp. 921–928, 2019.

[6] J. P. Klein, “Small Sample Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators,” Scandinavian Journal of Statistics, vol. 18, no. 4, pp. 333–340, 1991.

[7] “Lecture 2 estimating the survival function - UCSD mathematics,” Lecture 2 ESTIMATING THE SURVIVAL FUNCTION. [Online]. Available: https://www.math.ucsd.edu/ rxu/math284/slect2.pdf. [Accessed: 26-Nov-2021].

[8] M. Falcaro and J. R. Carpenter, “Correcting bias due to missing stage data in the non-parametric estimation of stage-specific net survival for colorectal cancer using multiple imputation,” Cancer epidemiology, vol. 48, pp. 16–21, 2017, doi: 10.1016/j.canep.2017.02.005.

[9] M. Jay and R. A. Betensky, “Displaying survival of patient groups defined by covariate paths: Extensions of the Kaplan-Meier estimator,” Statistics in Medicine, vol. 40, no. 8, pp. 2024–2036, Feb. 2021.

[10] M. K. Goel, P. Khanna, and J. Kishore, “Understanding survival analysis: Kaplan-Meier estimate,” International journal of Ayurveda research, Oct-2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/. [Accessed: 25-Nov-2021].

[11] M. Wakin, ”Lecture Notes, Set E: Continuous Random Variables,” presented to EENG 311, Colorado School of Mines, Golden, CO, USA, Oct. 6, 2021. [PowerPointslides]. Accessed on: Nov. 29, 2021.

[12] M. Wakin, ”Lecture Notes, Set H: Parameter Estimation,” presented to EENG 311, Colorado School of Mines, Golden, CO, USA, Dec. 3, 2021. [PowerPointslides]. Accessed on: Dec. 9, 2021.

[13] Saranya and Karthikeyan, “A Comparison study of Kaplan Meier and Nelson Aalen Methods in Survival Analysis,” INTERNATIONAL JOURNAL FOR RESEARCH IN EMERGING SCIENCE AND TECHNOLOGY, vol. 2, no. 11, pp. 34–38, Nov. 2015.

[14] S. Shapiro, “Evidence on screening for breast cancer from a randomized trial,” Cancer, vol. 39, no. 6, pp. 2772–2782, 1977.

[15] “Survival Distributions, Hazard Functions, Cumulative Hazards.” [Online]. Available: https://web.stanford.edu/ lutian/coursepdf/unit1.pdf. [Accessed: 01-Dec-2021].

[16] T. S. Robinson, “10 fundamental theorems for econometrics,” 30-Sep-2020. [Online]. Available: https://bookdown.org/ts_robinson1994/10_fundamental_theorems_for_econometrics/. [Accessed: 27-Nov-2021].

[17] Van Belle, P. Neven, V. Harvey, S. Van Huffel, J. A. . Suykens, and S. Boyd, “Risk group detection and survival function estimation for interval coded survival methods,” Neurocomputing (Amsterdam), vol. 112, pp. 200–210, 2013, doi: 10.1016/j.neucom.2012.12.049.

[18] V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. . Suykens, “Support vector methods for survival analysis: a comparison between ranking and regression approaches,” Artificial intelligence in medicine, vol. 53, no. 2, pp. 107–118, 2011, doi: 10.1016/j.artmed.2011.06.006.

[19] Willems, SJW, et al. “Correcting for Dependent Censoring in Routine Outcome Monitoring Data by Applying the Inverse Probability Censoring Weighted Estimator.” Statistical Methods in Medical Research, vol. 27, no. 2, Feb. 2018, pp. 323–335, doi:10.1177/0962280216628900.

[20] Y. Shen and J. Cai, “Maximum of the Weighted Kaplan-Meier Tests with Application to Cancer Prevention and Screening Trials,” Biometrics, vol. 57, no. 3, pp. 837–843, 2001, doi: 10.1111/j.0006-341X.2001.00837.x.