Hindawi
International Journal of Mathematics and Mathematical Sciences
Volume 2021, Article ID 9307475, 16 pages
https://doi.org/10.1155/2021/9307475

Review Article

Handling Censoring and Censored Data in Survival Analysis: A Standalone Systematic Literature Review

Anthony Joe Turkson[1], **Francis Ayiah-Mensah**[1] **and Vivian Nimoh**[2]

[1]*Takoradi Technical University, Mathematics, Statistics and Actuarial Science Department, Sekondi-Takoradi, Ghana*
[2]*Holy Child College of Education, Mathematics and ICT Department, Sekondi-Takoradi, Ghana*

Correspondence should be addressed to Anthony Joe Turkson; anthony.turkson@ttu.edu.gh

Received 16 August 2021; Revised 13 September 2021; Accepted 15 September 2021; Published 24 September 2021

Academic Editor: Niansheng Tang

Copyright © 2021 Anthony Joe Turkson et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The study recognized the worth of understanding the how's of handling censoring and censored data in survival analysis and the potential biases it might cause if researchers fail to identify and handle the concepts with utmost care. We systematically reviewed the concepts of censoring and how researchers have handled censored data and brought all the ideas under one umbrella. The review was done on articles written in the English language spanning from the late fifties to the present time. We googled through NCBI, PubMed, Google scholar and other websites and identified theories and publications on the research topic. Revelation was that censoring has the potential of biasing results and reducing the statistical power of analyses if not handled with the appropriate techniques it requires. We also found that, besides the four main approaches (complete-data analysis method; imputation approach; dichotomizing the data; the likelihood-based approach) to handling censored data, there were several other innovative approaches to handling censored data. These methods include censored network estimation; conditional mean imputation method; inverse probability of censoring weighting; maximum likelihood estimation; Buckley-Janes least squares algorithm; simple multiple imputation strategy; filter algorithm; Bayesian framework; β-substitution method; search-and-score-hill-climbing algorithm and constraint-based conditional independence algorithm; frequentist; Markov chain Monte Carlo for imputed data; quantile regression; random effects hierarchical Cox proportional hazards; Lin's Concordance Correlation Coefficient; classical maximum likelihood estimate. We infer that the presence of incomplete information about subjects does not necessarily mean that such information must be discarded, rather they must be incorporated into the study for they might carry certain relevant information that holds the key to the understanding of the research. We anticipate that through this review, researchers will develop a deeper understanding of this concept in survival analysis and select the appropriate statistical procedures for such studies devoid of biases.

1. Introduction

The aim of all researches is to advance, refine, and expand knowledge. It is also to establish facts and arrive at new conclusions using systematic inquiries and well documented methods. Longitudinal studies are a useful tool in most researches, offering numerous advantages over cross-sectional studies. By following the same subjects over a time period, longitudinal studies can be used to investigate the effects of predictors on disease and disease progression. The massive abundance of studies in biomedicine, sociology, demography, criminology, engineering, and economics, calls for a new dimension in studies on censoring and censored data. Survival analysis uses longitudinal data in analysis [1]. Survival analysis is a class of statistical methods for studying the occurrence and timing of events. Simply put survival analysis models time to failure or time to occurrence of an event. An event is a qualitative change that occurs at a given point in time to an individual, organization, entity, or society, although more than one event may be considered in the analysis, the assumption here is that only one event is of designated importance. Doing survival analysis requires that the researcher provides data for individuals for a time period prior to the occurrence of a given event [1, 2]. The survival

analysis procedure requires that researchers examine the effect of changes in the covariates on the duration of time preceding the event as well as the probability that the event will occur [3]. One crucial objective of survival analysis is to obtain a measure of the effect that will describe the relationship between a predictor variable of interest and time to failure after adjusting for other variables identified in the study and included in the model. References [2, 4, 5] have posited that survival analysis techniques model the probability of a change in a dependent variable Y_t from an origin state j to a destination state k as a result of causal factors and that the duration of time between states is referred to as event time. According to them, event time is represented by a nonnegative random variable T that represents the duration of time until the dependent variable at time t_0 changes from state j to state k.

Definition 1. Given n observations $T_1, T_2, \ldots, T_n$ independently and identically distributed *(i.i.d)* with the same distribution as T, the empirical survival function S_n is defined for all values of t by the following:

$$S_n(t) = \frac{\text{number of observations} > y}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{(y,\infty)}(T_i), \tag{1}$$

and is an estimate of the survival function $S(t) = P(T > t)$. Definition 1 shows that for a fixed value of t, $S_n(t)$ is an average of n independent and identically distributed random variables, Z_i, say, where $Z = I_{(y,\infty)}(T)$. Therefore, $nS_n(t) = \sum_{i=1}^{n} Z_i$. Each of the Z_i has the Bernoulli probability distribution where $P(Z_i = 1) = P(T_i > t) = S(t)$ and $P(Z_i = 0) = P(T_i \leq t) = 1 - S(t)$. This common Bernoulli distribution, therefore, has a probability mass function

$$\frac{z \mid 0}{P(Z = z) \mid 1 - S(t)} \quad \frac{1}{S(t)}. \tag{2}$$

We note that this probability distribution has mean

$$E(Z_i) = S(t) \text{and variance } \operatorname{Var}(Z_i) = S(t)[1 - S(t)]. \tag{3}$$

This means that $nS_n(t)$ has a binomial distribution: $nS_n(t) \approx \operatorname{Bin}(n, S(t))$. We can use the formula to estimate the survival at a fixed value of t,

$$\begin{aligned} E[S_n(t)] &= \frac{1}{n}[nS(t)] = S(t), \\ \operatorname{Var}\left[S_n(t) = \frac{1}{n^2}\{nS(t)[1 - S(t)]\} = \frac{S(t)[1 - S(t)]}{n}\right. &. \end{aligned} \tag{4}$$

At a fixed point $t = t^*$, we estimate $S(t^*) = P(T > t^*)$, the probability of surviving beyond t^*; using the estimator $S_n(t^*)$, we note that $S_n(t^*)$ is unbiased for $S(t^*)$. To get the accuracy by which $S_n(t^*)$ estimates $S(t^*)$, it is appropriate to use an approximate confidence interval based on two standard errors:

$$S_n(t) \pm 2\sqrt{\frac{S_n(t^*)[1 - S_n(t^*)]}{n}}. \tag{5}$$

A distinctive feature that differentiates survival analysis from other statistical methods is that survival data are usually censored or incomplete. It is comforting to note that in survival analysis, only a part of the subjects will experience the event of interest during the course of the experiment; the other part will not experience the event of interest after the expiration of the study; survival times for this latter part of the subjects will therefore be unknown. What the researcher will know is that their survival times are in excess of the time for which the particular subject had been observed. While in other statistical concepts, these incomplete data will be disregarded, the story is different in survival analysis; this unknown or incomplete data is regarded as an important component in survival analysis that they are taken into account [6].

Over the years, the analysis of statistical experiments has been made more complicated with issues of censoring. With this issue of censoring comes a lot of analytical schemes that must be taken into consideration. The use of standard methods to analyzing censored data will generate results that, in a way, has some level of biases because some important information would be left out. Missing or incomplete data is a prevalent problem in many research studies. If this problem of incomplete data is not appropriately addressed, it can lead to biases and inefficient estimation that can impact the conclusions of the study.

This paper sought to bridge the gap between the known and the unknown issues about censoring. It does so by systematically reviewing the works of different authors and their understanding of the concept of censoring. We explained clearly the concept of censoring, lasing it with examples, reviewed some theorems about the concept, provided the varieties of censoring mechanisms, described common statistical methods used to analyze censored data, investigated the effects of different censoring assumptions in actual studies, provided some research findings on the various censoring mechanisms which were used in handling censoring and censored data, gave some practical cases of censoring from the health sciences, and finally provided direction for future research on censoring. The research question of interest was "how do researchers handle censoring and censored data"?

To achieve the objectives of the study, a systematic literature review was conducted in accordance with the standard structure as presented in Table 1 and Figure 1. Reference [7] has noted that a systematic literature review is a method that clearly identifies, evaluates, and synthesizes previously completed and recorded work produced by researchers, scholars, and practitioners. The systematic review process is driven by a clear research question, which guides the collection of literature, the extraction of data, and ultimately the analysis of the data. Several guidelines are available on how to conduct a systematic literature review. A quality literature review follows a four-phase process, including planning, literature selection, data extraction, and writing the review [7–11].

Table 1: A detailed account of the systematic literature review procedure.

Serial number	Characteristics of the research	Inclusion criteria
(1)	Outcome investigated	Censoring/censored data
(2)	Language	English
(3)	Area in which research was conducted	Across the globe
(4)	Time span	1958–2021
(5)	Procedure adopted	Searched through National Centre for Biotechnology Information (NCBI) Resources–PubMed Central and others using Google scholar
(6)	Type of censoring investigated	Interval, right and left-censoring random and double-censoring independent and non-informative censoring. type I and type II censoring
(7)	Research design	Mata-analysis
(8)	Study design	Qualitative- using organized method in locating, assembling and evaluating body of literature on a particular topic

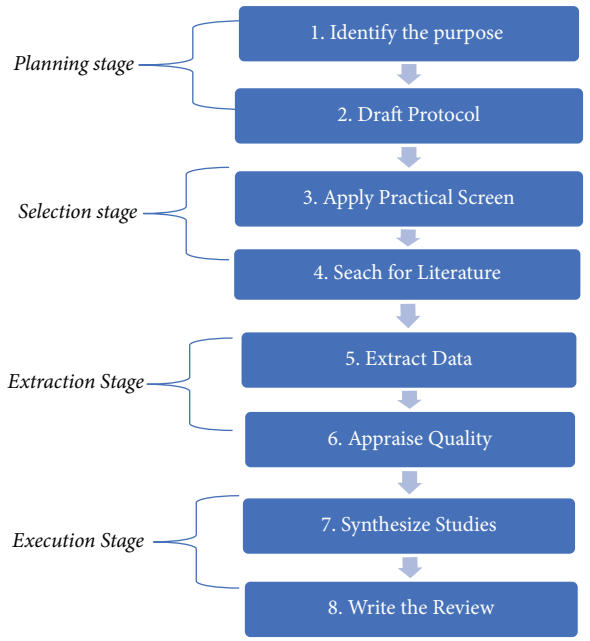

Figure 1: A systematic guide to literature review development.

1.1. The Concept of Censoring. The first and most crucial issue which survival analysis addresses is censoring. Censoring occurs when the event of interest is not observed for some subjects before the study is terminated. It occurs when the researcher has partial information about the subjects' survival times but is not privy to the exact survival times. For instance, if a researcher is observing 50 newly enlisted people who have had contacts with persons diagnosed with the COVID-19 pandemic during a 14-day quarantine and observation period with the view of finding out how long it takes them to show signs of the virus, it is possible that some persons might abscond from the place, other persons might die along the way due to unrelated diseases, and still, few of the people may not show any sign of the virus after the 14-day observation period, yet show signs after the quarantine period. Survival times for all these three cases may not be known to the researcher. They will only know that the duration is at least that much but would not know exactly what the survival times will be.

In a survival analysis study, the variable "Time" records two different things, for those subjects who obtained the event, it records the actual survival time, but for those who did not obtain the event of interest or those who were lost to follow-up, their progress could no longer be followed. Moreover, their true "failure" time will be longer than could be observed. Such subjects could be referred to as right-censored because, on a timeline, their true lifetimes are to the right of their observed censor times. In this case, time indicates the length of follow-up, which is a partial or incomplete observation of survival times. Survival analysis models use methods of estimation that incorporate information from censored and uncensored observation to provide consistent parameter estimates [1, 2].

1.2. Varieties of Censoring Mechanisms

1.2.1. Left-Censoring, Interval-Censoring, and Right-Censoring. A time X associated with a specific subject in a study is considered to be left-censored if it is less than the censoring time a where a is the left-censoring time. In other words, for left-censoring to occur, the event of interest must occur for the subject before that person is observed in the study at time a ($T < a$). For such subjects, we know that they have experienced the event sometime before time a, but the exact time is not known. The exact time will be known if and only if X is greater than or equal to a. The data from a left-censored sampling scheme can be represented by pairs of random variables (T, δ), where T is equal to X, if the lifetime is observed and δ indicates whether the lifetime corresponds to an observed event ($\delta = 1$) or is censored ($\delta = 0$), for left-censoring $T = \max(X_i, a)$.

When a specific subject is followed for a while, gets lost to follow-up for a while, and returns and continues being studied, such a subject is said to be interval-censored. ($a < T < b$). In interval-censoring, the observed data consists of intervals $I_1, I_2, \ldots, I_n$ where for each $i = 1, 2, \ldots, n$, the i^{th} response lies in the interval I_i. In this case, an uncensored observation of an observed death corresponds to an observed interval consisting of a single point. Suppose we have the situation where we have performed a test on a subject at some point in time ($t1$) and the subject tested

negative. But then at a time point further on ($t2$), the subject tested positive. In this scenario, we know the subject was exposed to the virus sometime between $t1$ and $t2$, but we do not know the exact timing of the exposure. For example, if in a clinical trial, the time to remission has been assessed, then if the i^{th} patient is in remission at the 8^{th} week after the trail, but was absent for future check-ups, and resurfaces and was out of remission on the 11^{th} week, then $I_i = [8, 11)$ is the i^{th} patients censoring interval or length of remission [12].

For a specific subject under study, if we assume that there is a time X and a censoring time b, the $X's$ are independent and identically distributed with probability function $f(t)$ and survival function $S(t)$. The exact lifetime X of a subject will be known, if and only if X is less than or equal to b; if X is greater than b, the subject is a survivor and his event time is censored at b. The data from this experiment can be represented by pairs of random variables (T, δ), where δ indicates whether the lifetime corresponds to an event $(\delta = 1)$ or is censored $(\delta = 0)$, and T is equal to X if the lifetime is observed and b if it is censored. For a right-censoring $T = \min(X_i, b)$, where T is some time variable and a and b some points in time [12].

Definition 2. The survival variables $Y_1, Y_2, \ldots, Y_n$ are right-censored by fixed constant $t_1, t_2, \ldots, t_n$, if the observed sample consists of the ordered pairs (Z_i, δ_i) for $i = 1, 2, 3, \ldots, n$. For each i: $Z_i = \min\{Y_i, t_i\}$,

$$\delta = \begin{cases} 1, & \text{if } Y_i \le t_i \quad \text{uncensored} \\ 0, & \text{if } Y_i > t_i \quad \text{censored} \end{cases}, \tag{6}$$

where t_i is the fixed censor time and δ is the censor indicator for T_i.

For left-censored data, the observed times are $Z_i = \max\{Y_i, l_i\}$, where l_i is the left censor time associated with Y_i. For left-censored data, $-Z_i = \min\{-Y_i, -l_i\}$.

It follows from the above that left-censoring is a special case of right-censoring with the time axis reversed. It is because of this phenomenon that there have been few specialist techniques developed explicitly for left-censored data [12].

1.2.2. Reasons for Right-Censoring. It comes about in the following ways: Study ends without subject experiencing the event; the subject is lost to follow-up within the study period; subject deliberately withdraws the treatment variable; the subject is obliged to withdraw from the treatment due to reasons beyond their control and subject withdraws from the study due to another reason (i.e., death, if death is not the event of interest).

Reference [13] has observed in a clinical trial that possible causes of patient withdrawal or dropping out of the study included death, adverse reactions, unpleasant study procedures, lack of improvement, early recovery, and other factors related or unrelated to trial procedure and treatments. In other cases, some data may not be collectible, observable, or available for some study subjects. Data can also be missing by the design of the study as a result of resource constraints [14].

1.2.3. Type I and Type II Censoring. Type I censoring occurs when a study is designed to end at a fixed period T_i of time which was fixed by the researcher. At the end of the study period, any subject that did not experience the event is censored. In type I censoring, the number of uncensored observations is a random variable.

In Type II censoring, the time may be left open at the beginning. The study is allowed to run until a prespecified fraction r/n of the n items has "failed." Let $T_1, T_2, \ldots, T_n$ denote the ordered values of the random sample $T_1, T_2, \ldots, T_n$ The observation terminates after the r^{th} failure time occurs, so we will only observe the r^{th} smallest observations in a random sample of n items.

Alternatively, let's suppose that for a given sample $T_1, T_2, \ldots, T_n$ of size n, only the first $r < n$ lifetimes are observed. The value of r is fixed before the survival data are seen. This means that the observed data consists of the smallest r observations. In terms of random variables, this may be expressed using other statistics. From the possible responses $T_1, T_2, \ldots, T_n$, we observe only the first r ranked responses $T_1, T_2, \ldots, T_n$. This means that

$$t_{(1)} \le t_{(2)} \le \cdots \le t_{(r-1)} \le t_{(r)}. \tag{7}$$

1.2.4. Random, Double, Independent, and Noninformative Censoring. In random censoring, the total period of observation is fixed, but subjects enter into the study at different points in time. Some subjects experience the events of interest. Others do not, and some are lost to follow-ups. Others will still be alive at the end of the study period. In random censoring, the censored subjects do not all have the same censoring time. Random censoring can also be produced when there is a single termination time, but entry times vary randomly across the subjects [14].

Double-censoring occurs as a result of a combination of left and right-censoring. In this case, we note that $Z_i = \max\{\min[T_i, t_i], l_i\}$, where l_i and t_i are, respectively, the left and right-censoring times associated with T_i, $l_i < t_i$, in this case, T_i is only observed if it falls in a window of observations $[l_i, t_i]$. Otherwise, one of the endpoints of the windows is observed and the other endpoint of the window probably remains undisclosed. We should also note that double-censoring is not the same as interval-censoring [4, 15].

Independent censoring occurs whenever there exists a subgroup of interest such that the subjects who are censored at time t are representative of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience. In other words, items may not be censored (withdrawn) because they have a higher or lower risk than the average. Reference [16] have noted that in practice, this assumption of independent censoring deserves special scrutiny, that is to say, the Kaplan–Meier estimator may overestimate the survival function of T, if the survival time and the censoring time are positively correlated and underestimate the survival function if the survival and censoring times are negatively correlated. Independent censoring may not hold for all situations but under some

dependence conditions we can still use the likelihood function [17, 18].

With *noninformative censoring*, participants who drop out of the study must not do so due to reasons unrelated to the study. Noninformative censoring occurs if the distribution of survival times (T) provides no information about the distribution of censorship times C and vice versa. That is, the reason why the time of the event was not observed was entirely unrelated to the outcome under study. The study simply ended while the observed subjects were alive. One consequence of noninformative censoring is that the underlying probabilities of obtaining the event of interest are the same for both censored and uncensored observations. The difference between the two is that the survival time for obtaining the event of interest for the censored observations is not known. Informative censoring occurs when subjects are lost to follow-up due to reasons related to the study [17–20].

1.2.5. Censoring Status. The censoring status is a dichotomous random variable $\delta = (1.0)$. When a subject obtains the event of interest, we denote the censoring status by 1. If the subject fails to obtain the event of interest, we denote the status by 0.

$$\delta = \left\{ \begin{array}{ll} 1, & \text{if failure} \\ 0, & \text{if censored} \end{array} \right\}. \tag{8}$$

In Figure 2, the bullets at the end of lines representing Anthony, Bernard, Daniel, and Edward mean that each of them obtained the event of interest. That is to say, they were uncensored. The symbol O at the end of the line "Charles" means that he was censored. Anthony got the event before the end of the study. He was, therefore, uncensored. Bernard got the event before the observation period. Therefore, he was left-censored. Charles got enrolled in the study; at the start of the study, he was observed for some time; he took a break and was nowhere to be found (lost to follow up); he resurfaced after some time to continue with the study; he did not get the event of interest before the observation period ended, his status was, therefore, interval-censored. Finally, Edward got the event of interest after the observation period had come to an end. His status was therefore right-censored [19].

Theorem 1. *Under type I right-censoring with fixed censor times, the joint likelihood, $L(\varphi)$, of the observed data (Z_i, δ_i), $i = 1, 2, 3, \ldots, n$ is given by $L(\varphi) = k \prod_{i=1}^{n} f(z_i)^{\delta_i} S(z_i)^{1-\delta_i}$, where k is a constant.*

By a type I censoring design we mean a study in which every subject is under observation for a specified period C_0 or until failure. A slightly more complicated type I censoring design is one in which each of the subjects have their own *fixed* censoring time C_1, instead of a common censoring time C_0. In this study design, the likelihood function for each subject can be represented by one of the following two probabilities: the probability that the event occurred in a small interval including time t_i [denoted by $f_i(t_1)$ or the probability that the subject did not have the event at C_1 [denoted by $S_i(C_1)$].

Proof. We start the proof by considering a joint density function $f(z, \delta)$ of Z and δ, where $Z = \min\{Y, t\}$ for censoring indicator δ and t fixed. When $\delta = 0$, then conditionally $Z = t$ with probability $P = 1$, so that

$$\begin{aligned} f(z, 0) &= P(Z = t \mid \delta = 0) P(\delta = 0) \\ &= P(\delta = 0) = P(Y > t) = S(t). \end{aligned} \tag{9}$$

When $\delta = 1$, then $Z < t$ and $f(z|\delta = 1) = f(z)/F(t)$, is the conditional density used in the calculation of mean residual lifetime, using the conditional argument

$$f(z, 1) = f(z \mid \delta = 1) P(\delta = 1) = \left[\frac{f(z)}{F(t)} \right] F(t) = f(z). \tag{10}$$

Combining (9) and (10), we shall obtain the following:

$$f(z, \delta) = \left\{ \begin{array}{ll} f(z) & \text{if } \delta = 1 \\ S(t) & \text{if } \delta = 0 \end{array} \right\}. \tag{11}$$

Equation (11) can explicitly be written as follows:

$$f(z, \delta) = f(z_i)^{\delta} S(z_i)^{1-\delta}. \tag{12}$$

The joint distribution in (12) above is a constant multiple of the product of $f(z_i, \delta_i)$, $i = 1, 2, \ldots, n$.

Equation (12) can be generalized to accommodate other types of censoring, such as interval-censoring. If we assume that the interval-censoring mechanism operates independently of the observed lifetimes, then it follows that if $\delta = 0$ were represent an interval-censored observation $[a, b)$, then the contribution to the likelihood may be determined by the following:

$$\begin{aligned} f(z, 0) &= P(Z = t \mid \delta = 0) P(\delta = 0) = P(\delta = 0) \\ &= P(a \le Y < b) = S(a) - S(b). \end{aligned} \tag{13}$$

This means that terms of this form may be included in the likelihood in Theorem 1 for interval-censored observations.

The likelihood for various types of censoring schemes may be written by incorporating the following components:

Observed death at y $f(y)$

Right-censored observations at t $S(t)$

Left-censored observation at a $1 - S(a)$

Interval-censored observation $S(a) - S(b)$

The generalized likelihood function becomes as follows:

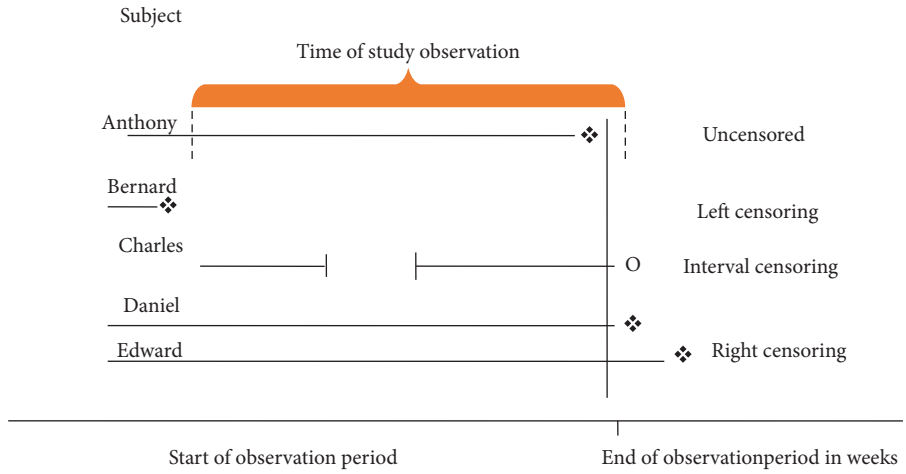

Figure 2: A study on five males illustrative of the three types of censoring.

$$L(\varphi) = k\prod_{D} f(y)\prod_{R} S(t)\prod_{L} 1 - S(a)\prod_{I}[S(a) - S(b), \tag{14}$$

where k is a constant; D, the set of observed data; R, the set of right-censored observations; L, the set of left-censored observations, and I the set of interval-censored observations.

The likelihood of random right-censored data is constructed in this manner; it is constructed so that φ will be estimated by maximum likelihood; it is observed that the ordered pairs are given by

(Z_i, δ_i), for$i = 1, 2, 3, \ldots, n$. For each i: $Z_i = \min\{Y_i, t_i\}$,

$$\delta_i = \left\{\begin{array}{ll} 1, & \text{if } Y_i \le t_i \quad \text{uncensored} \\ 0, & \text{if } Y_i > t_i \quad \text{censored} \end{array}\right\}, \tag{15}$$

where t_i is the fixed *censor time* and δ is the *censor indicator* for Y_i. The likelihood factorizes into two parts. One relates to the censor times and the other the survival or lifetimes. □

Theorem 2. *The joint likelihood $L(\varphi)$ of the observed data (Z_i, δ_i), $i = 1, 2, 3, \ldots, n$, is given by the following:*

$$L(\varphi) = k\left[\prod_{i=1}^{n} G(z_i)^{\delta_i} g(z_i)^{1-\delta_i}\right]\left[\prod_{i=1}^{n} f(z_i)^{\delta_i} S(z_i)^{1-\delta_i}\right], \tag{16}$$

where k is a constant.

Proof. The joint density of Z *an d* δ, where $Z = \min\{Y, T\}$, for the indicator variable δ and T a random variable independent of Y.

For $\delta = 0$,

$$\begin{aligned} P(Z < z, \delta = 0) &= P(T < z, Y > T) \\ &= \int_0^z P(T < z, Y > T | T = t) g(t) \mathrm{d}t \\ &= \int_0^z S(t) g(t) \mathrm{d}t. \end{aligned} \tag{17}$$

Differentiating each side of the expression above with respect to z using the principles of calculus gives the following:

$$f(z, 0) = S(z) g(z). \tag{18}$$

For $\delta = 1$,

$$\begin{aligned} P(Z < z, \delta = 1) &= P(Y < z, Y \le T) \\ &= \int_0^z P(Y < z, Y \le T | Y = y) f(y) \mathrm{d}y \\ &= \int_0^z G(y) f(y) \mathrm{d}y. \end{aligned} \tag{19}$$

Differentiating each side of the expression above with respect to z, using the principles of calculus gives the following:

$$f(z, 1) = G(z) f(z). \tag{20}$$

Combining (18) (20), we shall obtain the following:

$$f(z, \delta) = [f(z) G(z)]^{\delta} [g(z) S(z)]^{1-\delta}. \tag{21}$$

Equation (20) is, by definition, a constant product of$f(z_i, \delta)$, $i = 1, 2, \ldots, n$, that is,

$$L(\varphi) = k\prod_{i=1}^{n} [f(z_i) G(z_i)]^{\delta_i} [g(z_i) S(z_i)]^{1-\delta_i}. \tag{22}$$

Equation (22) gives the required results upon rearrangements

The implications of Theorem 2 are enormous; firstly, it is important for estimating maximum likelihood involving random censor times. We note from the theorem that

$$\left[\prod_{i=1}^{n} G(z_i)^{\delta_i} g(z_i)^{1-\delta_i}\right] \tag{23}$$

involves the distribution of the censor times only. If there are no parameter estimates for G and g; that is to say, if G and g are independent of φ, then this term acts as a constant multiple in L, when L is maximized with respect to φ, which

takes us to the likelihood of Theorem 2 for fixed censor times. Secondly, if we regard the observed censor times as conditionally fixed at times $t_1, t_2, \ldots, t_n$, then the term in the bracket is factored out of the conditional likelihood, which takes us to the fixed likelihood in Theorem 2. The argument above favors the fixed censorship model, which appears in Theorem 2. It should also be noted that in reality the assumption that Y and T are independent is not likely to hold, let's consider a heart transplant study, where doctors get to know that age affects survival of patients, doctors will then admit younger transplant patients into the study; these young patients will then last throughout the duration of the study to be censored at the termination date; this makes the survival and the censoring time dependent.

The likelihoods constructed above are used primarily for analyzing parametric models; they also serve as the basis for determining the partial likelihoods used in semiparametric regression methods [17, 21, 22]. □

2. Methods for Analyzing Censored Data

Reference [18] has discussed four common statistical methods that could be used in analyzing censored data. The first complete-data analysis method is adopted when the researcher decides to ignore the censored observation and conducts the analyses on only the uncensored observations. This method has the advantage of simplicity. The disadvantages are many, and there is the loss of efficiency and estimation bias.

The second method to analyzing censored data is the imputation approach. This method is one of the popular methods for handling incomplete data but may not be appropriate for censored data. Additionally, the authors have posited that although this method seems acceptable, there were two underlying disadvantages: For right-censoring, it was noted that if there was an assumption that all censored cases failed (that is, got the event of interest) right after the time of censoring, then survival probabilities would be underestimated; on the other hand, if all censored cases never failed, then the survival probabilities would be overestimated. For interval-censoring, it was revealed that the inappropriateness of imputation was quite unclear. Another method was to assume that the failure time after censoring followed a specific model and estimate the model parameters in order to impute the residual survival time (time from censoring to failure). However, this method depends on the model assumptions, which are very difficult to check without information on survival after censoring. Many researchers use imputation techniques, especially right-point or mid-point imputation when the observations are interval-censored. This may be due to a lack of statistical software packages for analysis. It has been pointed out that both right-point and mid-point imputations may generate some biased results.

According to reference [9], the third method to analyzing censored data is dichotomizing the data. With this method, the problem of right-censoring and interval-censoring may be avoided if one analyzes the incidence of occurrence versus nonoccurrence of the event within a fixed period of time and disregards the survival times. In this case, the dichotomized data can be easily analyzed by the standard techniques for binary outcomes, such as contingency tables and logistic regression. However, there are some disadvantages associated with this method: It cannot distinguish between the loss to follow-up and end-of-study censoring; the variability in the timing of the event among those who had the event within the observation period cannot be modeled; no time-dependent covariates (such as age, smoking status or alcohol consumption status) can be used in modeling. The authors had further posited that the method of analyzing dichotomized data may be acceptable when the risk of failure was low, risk periods are long, and covariates are associated with preventing the event rather than prolonging the survival time. Such situations are common in many epidemiological studies.

The fourth and final method is the likelihood-based approach. This happens to be the most effective method of censoring problems. It uses methods of estimation that adjust for whether or not an individual observation is censored or not. Many of these methods could be viewed as maximizing the likelihood under certain model assumptions, including assumptions about the censoring mechanism. Likelihood-based methods include, for example, the Kaplan–Meier estimator of the survival function in a one-sample problem, the log-rank test for testing equality of two survival functions in a two-sample problem, and the Cox-regression and accelerated-failure-time models for analysis of time to event data with covariates. The main advantage of the likelihood-based method is that it utilizes all the information available. However, as in the other methods, assumptions about the censoring mechanism are still required.

3. Maximum Likelihood Estimation

This is by far the most important method for analyzing censored data. This method adjusts for whether the subjects observed were censored or not. It also utilizes all the information available. The statistical approaches that fall within the likelihood method includes Kaplan–Meier estimator of the survival function for a one sample problem, the log-rank test for testing the equality of two survival curves in a two-sample problem, and the Cox regression and accelerated failure time models for analyzing time to events with covariates. In constructing the likelihood for censored, it should be noted that the lifetimes and censoring times are independent; if they are not independent then special techniques must be used in constructing the likelihood function for censored. An observation corresponding to an exact event time provides information on the probability that the event occurs at this time which is approximately equal to the density function X at this time. For a right-censored observation, all we know is that the event time is larger than this time, so that the information is the survival function evaluated at the on-study time. Similarly, for a left-censored observation, all we know is that the event has already occurred, so that the contribution to the likelihood is the cumulative distribution evaluated at the on-study time.

For interval-censored data, we know only that the event occurred within the interval. Maximum likelihood estimation is used because it produces estimators that are consistent, asymptotically efficient, and asymptotically normal. If data is gathered for a sample of n individuals ($i = 1, 2, \ldots, n$), the data will consist of t_i, the time of the event (or if the observation is censored, the time of censoring), an indicator variable, ∂ representing the presence ($\partial = 0$) or absence ($\partial = 1$) of censoring, and a vector of covariates, In the absence of censored observations, the probability of observing the entire data is the product of the probabilities of observing the data for each specific individual. When the probability of each observation is represented by its probability density function, we obtain the likelihood function:

$$L = \prod_{i=1}^{n} f_i(t_i), \tag{24}$$

where L represents the probability of the entire data. If censoring is present, then the likelihood function becomes as follows:

$$L = \prod_{i=1}^{n} [f_i(t_i)]^{\delta_i} [S_i(t_i)]^{1-\delta_i}. \tag{25}$$

The likelihood function effectively combines uncensored and censored observations, in that if an individual is not censored, the probability of the event is $f_i(t)$, and if the individual is censored at t_i, the probability of the event is $S_i(t_i)$, the survivorship function evaluated at t_i. Taking the natural log of L, the objective is to maximize the expression.

$$\log(L) = \sum_{i=1}^{n} \delta_i \ln f_i(t_i) + \sum_{i=1}^{n} (1 - \delta_i) \ln S_i(t_i). \tag{26}$$

Once the appropriate distribution has been specified, the process reduces to using a numerative method such as the Newton-Raphson algorithm to solve for the parameters. Most computer soft wares use the maximum likelihood approach to fit regression models to survival data. Survival models can be usefully viewed as ordinary regression models in which the response variable is time. However, computing the likelihood function (needed for fitting parameters or making other kinds of inferences) is complicated by censoring. The likelihood function for a survival model, in the presence of censored data, can also be formulated as follows [23–25],

From calculus, the density function $f(t, x, \beta)$ is the ratio of the hazard function and the survivorship function; substituting into $l(\beta) = \prod_{i=1}^{n} \{[f(t_i, x_i\beta)]^{c_i} \cdot [S(t_i, x_i, \beta)]^{1-c_i}\}$ will yield the expression

$$h(t, x, \beta) = \frac{f(t, x, \beta)}{S(t, x, \beta)},$$

$$f(t, x, \beta) = h(t, x, \beta) S(t, x, \beta),$$

$$\begin{aligned} \text{Likelihood}(\beta) &= \prod_{t=1}^{n} \{h(t_i, x_i, \beta) S(t_i, x_i, \beta)^c S(t_i, x_i, \beta)^{1-c} \\ &= \prod_{t=1}^{n} \{h(t_i, x_i, \beta)^C S(t_i, x_i, \beta), \end{aligned}$$

$$\text{Log}\, L(\beta) = \sum_{i=1}^{n} C[h_o(t_i)] + C_i X_i \beta + e^{x\beta_i} \ln(S_o(t_i). \tag{27}$$

This requires that we maximize the likelihood function with respect to the parameter of interest β and the unspecified baseline hazard and survivorship functions [26–28].

Reference [23] proposed using an expression he called partial likelihood function that depends only on the parameter of interest; he posited that the resulting parameters from the partial likelihood function would have the same distributional properties as the full maximum likelihood estimators; mathematical proofs of this conjecture came later which was based on the counting process approach by Martingales as detailed in [29, 30]. The method of partial likelihood begins by assuming that there is a group of individuals, $R(t_{(i)})$, that are at risk of failure just before the occurrence of $t_{(i)}$. If only one failure occurs at $t_{(i)}$, the conditional probability that the failure occurs to individual i, given that individual i has a vector of covariates x_i is represented by the following:

$$\frac{h(t_{(i)}|x_{(i)})}{\sum_{i \in R(t_{(i)})} h(t_{(i)}|x_{(i)})} = \frac{h_0(t) e^{x_i\beta}}{\sum_{i \in R(t_{(i)})} h_0(t) e^{x_i\beta}} = \frac{e^{x_i\beta}}{\sum_{i \in R(t_{(i)})} e^{x_i\beta}}. \tag{28}$$

Equation (28) is the hazard function for individual i at a specific point in time, $t_{(i)}$ divided by the sum of the hazard functions for all individuals in the risk set just before the occurrence of time $t_{(i)}$. Because $h_0(t)$ is common to every term in the equation, it is eliminated. The partial likelihood function is obtained by taking the product of equation (28) overall $k$points in time such that

$$l_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{x_i\beta}}{\sum_{j \in R(t_i)} e^{x_j\beta}} \right]^{c_i}. \tag{29}$$

Equation (29) does not depend on $h_0(t)$ and can be maximized to provide an estimate of β that is consistent and asymptotically normally distributed on the assumption that there are no tied times and excludes terms when $c_i = 0$ [22].

4. Practical Cases of Censoring from the Health Sciences

Reference [18] has documented some examples of how censoring originates; one of such cases looked at the subject's National Health Insurance status and mortality. In the study, the time period of interest was the time from the start of follow-up to death. The goal was to examine the relationship between the status of the subject's insurance policy and the risk of mortality. The research question was whether this variable (mortality status) was affected or not by the insurance status of the adults. 250 adults who were older than 40 years and had reported their insurance status were followed for five years. At the termination of the study, there were some who had not obtained the event of interest and therefore were right-censored. The analysis was adjusted for other factors such as baseline age, gender, race, smoking status, alcohol consumption, obesity, and employment status.

In another development, a sample of 55 women and 45 men attending a smoking treatment Clinic were studied. A number of demographics were observed to help determine whether women and men revert to smoking for the same or different reasons. The length of follow-up was five months and the variable of interest was the time from the start of follow-up to reverting into smoking. After 5 months, 38 subjects were lost to follow-up (16 women and 22 men), and 62 subjects (39 women and 23 men) had reverted. We note especially that the number who were lost to follow-up were quite different between women and men.

5. Developments Made to Handle Censored Data

Although this problem of censoring presents an obstacle of distortion, reference [31] has advanced a multivariate survival analysis method that could handle multiple events with censoring. They have made it possible to measure a bivariate probability density function for a pair of events. They proposed a method called censored network estimation to discover partially correlated relationships and construct the corresponding network, which was composed of edges representing nonzero partial correlations on multiple censored events. To demonstrate its superior performance compared to conventional methods, they proposed a selection power for the partially correlated events which was evaluated in two types of networks with iterative simulation experiments. Reference [32] has identified methods for censored outcomes which have become abundant in the literature; according to the authors, these methods for censored covariates have received little attention and dealt only with the issue of limit-of-detection. They have noted in particular that, for randomly censored covariates, an often-used method was the inefficient Complete-Case Analysis (CCA) method which consisted in deleting censored observations in the data analysis. It was further noted that when censoring was not completely independent, the CCA method led to biased and spurious results. Additionally, they have noted that methods for missing covariate data, including type I and type II censoring as well as limit-of-detection, did not readily apply due to the fundamentally different nature of randomly censored covariates. They then developed a novel method for censored covariates using a conditional mean imputation method which was based on either Kaplan–Meier estimates or a Cox proportional hazards model to estimate the effects of these covariates on a time-to-event outcome. They have evaluated the performance of the proposed method through simulation studies and showed that the imputation method provided a good reduction in bias and improved statistical efficiency. Finally, they have illustrated the method using data from the Framingham Heart Study to assess the relationship between offspring and parental age of onset of cardiovascular events.

Reference [33] has worked on a general-purpose approach to account for right-censored outcomes using Inverse Probability of Censoring Weighting (IPCW). They illustrated how IPCW can easily be incorporated into a number of existing machine learning algorithms used to mine big health care data including Bayesian networks, k-nearest neighbors, decision trees, and generalized additive models. Furthermore, they showed that their approach leads to better calibrated predictions than the three ad hoc approaches when applied to predicting the 5-year risk of experiencing a cardiovascular adverse event. Reference [34] has noted that censoring due to a limit of detection or limit of quantification happens quite often in many medical studies and pointed out the conventional approaches to dealing with censoring when analyzing these data, which methods include the substitution method and the Complete Case Analysis (CCA). They clearly pointed out that the CCA and the substitution method usually led to biased estimates. They added that more recently, Maximum Likelihood Estimation (MLE) had been increasingly used. It was intimated that the MLE approach appeared to perform well in many situations. They proposed an MLE approach to estimate the association between two measurements in the presence of censoring in one or both quantities. The central idea was to use a copula function to join the marginal distributions of the two measurements. In various simulation studies, they showed that their approach outperforms existing conventional methods (CCA and substitution analyses). Furthermore, they proposed a straightforward MLE method to fit a multiple linear regression model in the presence of censoring in a covariate or both the covariate and the response. Finally, they compared and discussed the performance of their method with multiple imputations and missing indicator model approaches.

Reference [35] looked at data below detection limits, that is to say, left-censored data, which were common in environmental microbiology. They then utilized simulated data sets informed by real-world enterovirus water data to evaluate methods that could be used for handling left-censored data. Four censoring degrees (low [10%], medium [35%], high [65%], and severe [90%]) and one real-life censoring example (97%) were simulated. For each of the data sets, five methods for handling left-censored data were applied:

(i) Substitution with a limit of detection LOD/$\sqrt{2}$

(ii) Lognormal Maximum Likelihood Estimation (MLE) to estimate mean and standard deviation

(iii) Kaplan–Meier estimation (KM)

(iv) Imputation method using MLE to estimate distribution parameters (MI method 1)

(v) Imputation from a uniform distribution (MI method 2)

The mean of each of the data sets was used to estimate enterovirus dose and infection risk. Root Mean Square Error (RMSE) and bias were used to compare estimated and known doses and infection risks. The following results were obtained; MI method 1 resulted in the lowest dose and infection risk. MI method 2 was the next overall best method. MI method 1 resulted in the least error. They concluded that if one was unsure of the distribution, MI method 2 would be the most preferred method. Reference [36] has proposed to reverse the Buckley-James least squares algorithm to handle left-censored data enhanced with a Lasso regularization to accommodate high-dimensional predictors. They presented a Lasso-regularized Buckley-James least squares method with both nonparametric imputations using Kaplan–Meier and parametric imputation based on the Gaussian distribution, which was typically assumed for HIV viral load data after logarithmic transformation. Their cross-validation for parameter-tuning was based on an appropriate loss function that took into account the different contributions of censored and uncensored observations. They specified how those techniques could be easily implemented using the R packages. They then did a comparative study of the Lasso-regularized Buckley-James least square method and simple imputation strategies to predict the response to antiretroviral therapy measured by HIV viral load. Reference [37] considered interval-censoring as a missing data problem and implemented a simple multiple imputation strategy that allowed flexible sensitivity analyses with respect to the shape of the censoring distribution. This they did to allow the identification of appropriate parametric models. They derived a χ^2-goodness-of-fit test which was supplemented with diagnostic plots. Finally, they obtained uncertainty estimates for mean survival times via a simulation strategy. They validated the statistical efficiency of the proposed method for varying lengths of intervals in a simulation study and compared it with simpler alternatives. Reference [38] has presented the equations that underlie the proportional subdistribution hazards model to highlight the way in which the censoring distribution was included in its estimation via risk set weights. They achieved that through simulating competing risk data under a proportional subdistribution hazards model with different patterns of censoring. In their work, they examined the properties of the estimates from such a model when the censoring distribution was misspecified. Reference [39] observed that censored data occur commonly in trial-structured behavioral experiments and many other forms of longitudinal data. They noted in particular that censored data have the tendency of biasing and reducing the statistical power in subsequent analyses. They acknowledged that the principled approaches for dealing with censored data, which included data imputation and methods based on the complete data's likelihood, work well for estimating fixed features of statistical models but could not be extended to dynamic measures, such as serial estimates of an underlying latent variable over time. Due to this limitation, they proposed an approach to the censored-data problem for dynamic behavioral signals. This was achieved by a state-space modeling framework with a censored observation process at the trial timescale. They then developed a filter algorithm to compute the posterior distribution of the state process using the available data. They showed that special cases of this framework could incorporate the three most common approaches to censored observations: ignoring trials with censored data, imputing the censored data values, or using the full information available in the data likelihood. Finally, they derived a computationally efficient approximate Gaussian filter that was similar in structure to a Kalman filter but efficiently accounted for censored data.

Reference [40] recognized the presence of Left-censoring in salivary bioscience data which occurs when salivary analyte determinations fell below the lower limit of an assay's measurement range. They noted that conventional statistical approaches for addressing censored values (i.e., recoding as missing, substituting, or extrapolating values) may introduce systematic bias. Furthermore, they elucidated that specialized censored data statistical approaches (like Maximum Likelihood Estimation, Regression on ordered Statistics, Kaplan–Meier, and general Tobit regression) were available. They noted that even though these methods were available, they were rarely being implemented in bio-behavioral studies that examine salivary biomeasures and their application to salivary data analysis. Reasons they adduced to this apparent nonuse of these approaches may be due to their sensitivity to skewed data distributions, outliers, and sample size. In their study to address the challenges, they compared descriptive statistics, correlation coefficients, and regression parameter estimates generated via conventional and specialized censored data approaches using salivary C-reactive protein data and assessed the differences in statistical estimates across approach and across two levels of censoring (9% and 15%). They examined the sensitivity of their results to sample size. It was revealed in their study that the results were similar across conventional and censored data approaches. They noted in particular that the implementation of specialized censored data approaches was more efficient (i.e., required little manipulations to the raw analyte data) and appropriate. Based on their findings, they outlined preliminary recommendations to enable investigators to more efficiently and effectively reduce statistical bias when working with left-censored salivary bio-measure data.

Reference [41] conducted a study on HIV vaccines and noted that longitudinal immune response biomarker data were often left-censored due to lower limits of quantification of the employed immunological assays. They were of the view that censoring information was important for predicting HIV infection-the failure event of interest. In their study, they proposed two approaches to addressing left-

censoring in longitudinal data: one that made no distributional assumptions for the censored data–treating left-censored values as a "point mass" subgroup–and the other which made a distributional assumption for a subset of the censored data but not for the remaining subset. These two approaches were developed to handle censoring for joint modeling of longitudinal and survival data via a Cox proportional hazards model fitted by h-likelihood. Their work was evaluated via simulation with HIV vaccine trial data set, and they found that longitudinal characteristics of the immune response biomarkers were highly associated with the risk of HIV infection.

Reference [42] penned down a study aimed at evaluating the impact of different proportions (i.e., 20%, 40%, 60% and 80%) of censored (CEN) or penalized (PEN) data in the prediction of breeding values (EBVs), genetic parameters, and computational efficiency for two longevity indicators (i.e., traditional and functional longevity; TL and FL, respectively). Three different criteria were proposed for PEN: (1) assumed that all cows with censored records were culled one year after their last reported calving; (2) assumed that cows with censored records older than nine years were culled one year after their last reported calving, while censored (missing) records were kept for cows younger than nine years; (3) assumed that cows with censored records older than nine years were culled one year after their last reported calving, while cows younger than nine years were culled two years after their last reported calving. They performed all analyses using random regression models based on fourth-order Legendre orthogonal polynomials. The proportion of commonly selected animals and EBV correlations were calculated between the complete dataset (i.e., without censored or penalized data; COM) and all simulated proportions of CEN or PEN. The computational efficiency was evaluated based on the total computing time taken by each scenario to complete 150,000 Bayesian iterations. The results were revealing.

Reference [42] has noted that classical statistical methods used for analyzing exposure data with values below the detection limits were well described in the occupational hygiene literature but revealed that evaluation of such data using the Bayesian approach was currently lacking. They proceeded to describe a Bayesian framework for analyzing censored data and presented the results of a simulation study conducted to compare the β-substitution method with a Bayesian method for exposure datasets drawn from lognormal distributions and mixed lognormal distributions with varying sample sizes, geometric standard deviations (GSDs) and censoring for single and multiple limits of detection. For each set of factors, estimates for the arithmetic mean (AM), geometric mean, GSD, and the 95th percentile ($X_{0.95}$) of the exposure distribution were obtained. They evaluated the performance of each method using relative bias, and the root mean squared error (rMSE), and coverage (the proportion of the computed 95% uncertainty intervals containing the true value). The results revealed that the Bayesian method using noninformative priors and the β-substitution method were generally comparable in bias and rMSE when estimating the AM and GM. For the GSD and the 95th percentile, the Bayesian method with noninformative priors was more biased and had a higher rMSE than the β-substitution method, and they noted that the use of more informative priors generally improved the Bayesian method's performance, making both the bias and the rMSE more comparable to the β-substitution method. It was indicated that the advantage of the Bayesian method was that it provided estimates of uncertainty for these parameters of interest and good coverage, whereas the β-substitution method only provided estimates of uncertainty for the AM, and coverage was not as consistent as the Bayesian method. It was concluded that the selection of one or the other method depends on the needs of the practitioner, the availability of prior information, and the distribution characteristics of the measurement data. They suggested to researchers to resort to the use of Bayesian methods if the practitioner has the computational resources and prior information, as the method would generally provide accurate estimates and also provide the distributions of all of the parameters, which could be useful for making decisions in some applications.

Reference [43] has proposed a method of handling censored survival data by assigning distributions of outcomes to shortly observed censored instances. Their work was an extension of the work of [44], who applied their learning technique to two well-known procedures for learning Bayesian networks: a search-and-score hill-climbing algorithm and a constraint-based conditional independence algorithm. They divided the data into three groups, as suggested by [45]: (1) instances for which the event occurred at any time were labeled positive; (2) instances censored after a certain critical time point $T*$ were labeled negative (event-free); (3) instances event was censored before time point $T*$ were doubled, split into both possible outcomes, and then assigned an estimated probability of an outcome based on the Kaplan–Meier method. The method was thoroughly tested in a simulation study and on the publicly available clinical dataset GBSG2. They compared it to learning Bayesian networks by treating censored instances as event-free and to Cox regression. Their results on model performance suggested that the weighting approach performed best when dealing with intermediate censoring. The result revealed that there was no significant difference between the model structures learned using either the weighting approach or by treating censored instances as event-free, regardless of censoring.

Reference [46] has noted that Interval-censoring appears when the event of interest is only known to have occurred within a random time interval. They adopted a procedure for performing estimation and hypothesis testing for interval-censored data. They distinguished between frequentist and Bayesian approaches. Computational aspects for every proposed method were described and solutions obtained with S-Plus.

As an improvement to the midpoint censoring interval, reference [47] proposed the Markov Chain Monte Carlo for imputed data MCMCid; in this method, the researcher sampled a value of each interval-censored origin from an estimated distribution of origins. In their approach, G

denoted the distribution of origins y, the estimate of G that is, $\widehat{G}$ was obtained parametrically using MLE based on a known distribution (that is, Weibull and Log normal) or nonparametrically using Turnbulls self-consisting algorithm, they showed that to perform an analysis for doubly censored survival data using MCMCid for each subject $i = 1, 2, \ldots, n$. A value of y_i denoted by $\widehat{y}$, is randomly sampled from $\widehat{a}$, conditional on the interval $[l_i, \mu_i]$ within which y_i falls. The doubly censored data set C and imputed origins $\widehat{y} = (\widehat{y}_1 \widehat{y}_2 \ldots \widehat{y}_n)^T$ are then mapped to a right-censored data set $D(C, \widehat{y}) = \left\{[\widehat{t}_i, \delta_i, x_i, z_i\right\}: i = 1, 2, \ldots, n]$, where $\widehat{t} = e_i - \widehat{y}_i$. The hierarchical Cox proportional hazards model could then is fit to the right-censored data $D(C, \widehat{y})$ using MCMC methods. They noted that even though the MCMC approach was straightforward to implement and to understand, it underestimated the variability of parameter estimates because of the uncertainty of imputed origins $\widehat{y}$, which was not incorporated. Reference [48] further discussed the imputation-embedded MCMC (*ie*MCMC) approach which was developed as an alternative to the MCMCid. For each MCMC iteration step m, $m = 1, 2, \ldots, M$, origins $y_1^m, y_2^m, \ldots, y_n^m$ were randomly sampled based on their distributions $\widehat{G}$ conditional on the intervals $[l_i, \mu_i]: i = 1, 2, \ldots, n$ within which they fall. A right-censored data set $D(C, y^m) = \{[t_i^m, \delta_i, x_i, z_i\}: i = 1, 2, \ldots, n]$, where $t^m = e_i - y_i^m$ was generated at each iteration step m. In the *ie*MCMC approach the origins y^m varied at each MCMC sampler iterations with a distribution based on $\widehat{G}$ but the estimate $\widehat{G}$, the uncertainty was estimating G by $\widehat{G}$, which was not taken into consideration although the uncertainty in y^m conditional on $\widehat{G}$ was considered.

Reference [48] has noted that dependent censoring occurs in many biomedical studies and poses considerable methodological challenges for survival analysis. They developed a new approach for analyzing dependently censored data by adopting quantile regression models. This was achieved by formulating covariate effects on the quantiles of the marginal distribution of the event time of interest. The strategy they adopted could accommodate a more dynamic relationship between covariates and survival time compared to traditional regression models in survival analysis, which usually assumed constant covariate effects. They then proposed estimation and inference procedures along with an efficient and stable algorithm. Again, they established a uniform consistency and weak convergence of the resulting estimators. This was done because an extensive simulation studies had the power to demonstrate good finite-sample performance of the proposed inferential procedures. They finally illustrated the practical utility of their method through an application to a multicenter clinical trial that compared warfarin and aspirin in treating symptomatic intracranial arterial stenosis.

Reference [49] has used quantile regression to analyze survival times, and he noted that the quantile method offered a valuable complement to traditional Cox proportional hazards modeling, which has been hampered by the lack of a conditional quantile estimator for censored data which was directly analogous to the Kaplan–Meier estimator and which applied under standard assumptions for censored regression models. He developed a recursively reweighted estimator of the regression quantile process, which was a direct generalization of the Kaplan–Meier estimator. He noted specifically that the asymptotic behavior was directly analogous to that of the Kaplan–Meier estimator, and its computation was essentially equivalent to current simplex methods for the quantile process in the uncensored case. Preliminary examples tested suggested a strong potential of these methods as a complement to the use of Cox models.

Reference [50] addressed two common statistical problems in pooling survival data from several studies. The first problem addressed was in respect of data that were doubly censored: Origin was interval censored while the endpoint event might be right-censored. Two approaches to incorporating the uncertainty of interval-censored origins were developed and then compared with more usual analyses using the imputation of a single fixed value for each origin. The second problem that was addressed was in respect of data collected from multiple studies and it was likely that heterogeneity existed among the study populations. To handle this data, a random-effects hierarchical Cox proportional hazards model was used. Reference [50] has indicated that the scientific problem that motivated their work was a pooled survival analysis of data sets from three studies meant to examine the effect of GB virus type C (GBV-C) coinfection on the survival of HIV-infected individuals. It was noted that the time of HIV infection was the origin and for each subject, this time was unknown, but was known to lie later than the last time at which the subject was known to be HIV negative, and earlier than the first time the subject was known to be HIV positive. They affirmed that their use of an approximate Bayesian approach using the partial likelihood as the likelihood was recommended because it more appropriately incorporates the uncertainty of interval-censored HIV infection times.

Reference [51] has underscored the fact that in many clinical biomarker studies, Lin's concordance correlation coefficient (CCC) was commonly used to assess the level of agreement of a biomarker measured under two different conditions. They noted, however, that, measurement of a specific biomarker typically could not provide accurate numerical values below the lower limit of detection (LLD) of the assay, which resulted in left-censored data. They alluded to the fact that most researchers discarded the data below the LLD or applied simple data imputation methods in the presence of left-censored data, such as replacing values below the LLD with a fixed number less than or equal to the LLD. Accordingly, they asserted that this method of discarding data below LLD was not statistically optimal because it often led to biased estimates and overestimates of the precision. They described a simple method using a bivariate normal distribution and applied SAS statistical software to arrive at the maximum likelihood (ML) estimate of the parameters. They constructed the estimate of the CCC and conducted a computer simulation study to investigate the statistical properties of the ML method versus the data deletion and simple data imputation method. They also contrasted the methods with real data using two urine

biomarkers, Interleukin 18 and Cystatin C. Their studies confirmed that the ML procedure was superior to the data deletion and simple data imputation procedures. In all of the simulated scenarios, the ML method yielded the smallest relative bias and the highest percentage of the 95% confidence intervals that included the true value of the CCC. In their first simulation scenario (sample size was 100 paired data points, 25% left-censoring for both members of the pair, true CCC was 0.238), the relative bias was −1.43% for the ML method, −40.97% for the data deletion method, and it ranged between −12.94% and −21.72% for the simple data imputation methods. Similarly, when the left-censoring for one of the members of the data pairs increased from 25% to 40%, the relative bias displayed the same pattern for all methods. It was concluded that when estimating the CCC from paired biomarker data in the presence of left-censored values, the ML method works better than data deletion and simple data imputation methods.

A laboratory study aimed at analyzing below detectable (BD) biomarkers lab values was studied. Four options for analyzing BD lab values were proposed.

> Option 1: If the % of BD was very high, the data was to be analyzed as a binary variable (detectable vs. undetectable)
>
> Option 2: If the % of BD was moderate) < 25%) discrete categories were to be created (tertile; quantiles; quintiles)
>
> Option 3: If the % of BD was rare (<5%) and values were highly skewed, then we need to replace BD values with LOD/2 where LOD = Limit of detection.
>
> Option 4: If the % of BD was rare (<5%) and values were less skewed, then we need to replace BD values with LOD/square root of 2.

It was also noted in their proposal that options 3 and 4 had the potential of introducing bias and reducing the power of the analysis. It was further noted that if the % of BD was extremely rare, it would have little effect on the results. For more complex BD values, it was proposed that

If the lab values were the outcome, we have to use Tobit regression to handle BD values (The BD values may first be normalized depending on the software used).

Secondly, we would have to use multiple imputation to replace BD values with imputed values.

Finally, we could use the propensity score weighting (modal probability of being BD, then incorporate the results in the modal using detectable values; https://khrc.ucsf.edu/analysis-below-detectable-biomaker-lab-values).

Reference [52] identified a new distribution for analyzing time-to-event data, which contained randomly censored data introduced by [53] known as generalized exponential (GE) distribution. Generalized exponential distribution could be used as an alternative to the well-known and used Weibull distribution in lifetime data analysis and reliability engineering.

The generalized exponential distribution, we are told, has the distribution, density, and survival functions, respectively, as follows:

$$\begin{aligned} F(t;\theta,p) &= 1-\exp\left(-\theta t^{p)}\right), \quad \theta, p>0, \\ f(t;\theta,p) &= p\theta\left[1-\exp\left(-\theta t^{p-1)}\right)\exp(-\theta t), \\ S(t;\theta,p) &= 1-\left[1-\exp\ (-\theta t)^{p}\right], \end{aligned} \tag{30}$$

where p is the shape parameter and θ the scale parameter. If the GE distribution with the shape p parameter and the scale parameter θ is denoted by GE(θ, p), then the two-parameter GE(θ, p) can be used quite effectively in analyzing many lifetime data and could assume the place of the two-parameter gamma and two-parameter Weibull distributions. The two-parameter GE(θ, p) could have increasing and decreasing failure rates depending on the shape parameter.

It was noted that studies that involved time-to-event or survival data analysis were focused on measuring the time-to-event of an outcome. It was further established that time-to-event could vary from time to either death or the occurrence of a clinical endpoint such as disease or the attainment of a biochemical marker.

Reference [52] also compared the classical maximum likelihood estimator to the proposed Bayesian estimators with two loss functions for the unknown parameters of the generalized exponential distribution for different sample sizes and parameter values.

The study revealed that the smallest standard error was Bayesian under the linear exponential loss function for both the scale and shape parameters. This happened under the Tierney and Kadane (T and K) numerical approximation procedure. This was followed by the Bayes estimator using the squared error loss function, again with the Tierney and Kadane (T and K) method. Reference [52] observed that the linear exponential loss function had the narrowest credible intervals with respect to the Tierney and Kadane approach as compared to the credible intervals of Bayes using Lindley and the confidence intervals obtained from the maximum likelihood estimator. This happened with a negative loss parameter, an indication of underestimation of the generalized exponential distribution parameters.

From the results and discussions obtained, reference [52] noted that it was evident that the Bayesian estimator under linear exponential loss function performed quite better than Bayes under squared error loss function and maximum likelihood estimator for estimating both the scale parameter and shape parameter. Reference [54] reviewed a comprehensive set of statistical methods for performing quantile regression with different types of survival data. The review covered various survival scenarios, including randomly censored data, data subject to left truncation or censoring, competing risks and semicompeting risks data, and recurrent events data. Two real examples were presented to illustrate the utility of quantile regression for practical survival data analyses.

6. Discussion and Conclusion

We have presented a review on censoring and censored data which are deemed to be unique concepts in survival analysis and longitudinal data analysis, in particular; we looked at

data below detection limit and data beyond the study period. Again, we looked at random, independent, type I and type II, informative, and noninformative censoring. The concept under review makes survival analysis appear difficult to some researchers whose knowledge about the concept is limited.

The study sought to pull ideas together on how researchers could handle censoring and censored data. Various studies on the topic have been documented. Our concern, therefore, was to bring these ideas under one umbrella. In an attempt to achieve our objectives, we chose a tried and tested approach of carrying out such reviews.

We adopted the systematic review method and searched through the National Centre for Biotechnology Information (NCBI) Resources–PubMed Central and others using Google scholar to extract articles and books that have been written on the subject matter and which relate to the research question-how do researchers handle censoring and censored data? Out of the over 1000 topics on the research question, we zeroed in on over forty (40) articles and books that carefully treated the topic of study.

We found in particular that censoring and censored data had a vast range of applications and have the potential of biasing results and reducing the statistical power of analyses if not handled with the appropriate techniques and tools it requires. We also found out that, besides the four main approaches to handling censored data, namely: complete data analysis method; imputation approach; dichotomizing the data and the likelihood-based approach, there were several other innovative approaches to handling censored data. These methods which have been tried and tested included Censored network estimation [30], the conditional mean imputation method [31], and inverse probability of censoring weighting [32]. Maximum likelihood estimation [33] Buckley-James's least squares algorithm [35]; simple multiple imputation strategy [36] and filter algorithm [38]. The majority of the authors modified and used the method of imputation. Other authors modified and extended the maximum likelihood estimation method [13].

Some authors concluded on the inefficiency of the complete case analysis [11, 13]. Additionally, the following methods for handling various types of censored data were investigated: The Bayesian framework for analyzing censored survival data [55]; the β-substitution method compared with the Bayesian method [42]; search-and-score-hill-climbing algorithm and constraint-based conditional independence algorithm on censored survival data [43, 44]; frequentist and Bayesian approaches to handling censored data and the Markov chain Monte Carlo for imputed data (MCMCid) [46, 47]; quantile regression [43, 48, 49]; random effects hierarchical Cox proportional hazards [50].

Lin's Concordance Correlation Coefficient (CCC) for assessing the level of agreement of a biomarker measured under two different conditions [51] and the classical maximum likelihood estimate [52].

In conclusion, we infer that the presence of incomplete information about subjects does not necessarily mean that such information must be discarded. Rather they must be incorporated into the study, for they might carry certain relevant information that might hold the key to the understanding of the research. We anticipate that through this review, researchers will develop a deeper understanding of this concept in survival analysis and select the appropriate statistical procedures for such studies devoid of biases

Data Availability

This is a review of literature that does not require any data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Anthony Joe Turkson performed conceptualization, gathering of data, and extraction of data. Francis Ayiah-Mensah developed the methodology and drafted and wrote the article. Vivian Nimoh reviewd and edited the article.

References

[1] P. D. Allison, *Survival Analysis Using the SAS System-A Practical Guide*, SAS Institute, Cary, NC, USA, 1995.

[2] P. D. Allison, *Event History Analysis*, Sage Publications, Beverly Hills, CA, USA, 1984.

[3] H.-P. Blossfeld, A. Hamerle, and K. U. Mayer, *Event History Analysis Statistical Theory and Application in the Social Sciences*, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1989.

[4] H.-P. Blossfeld and G. Rohwer, *Techniques of event history modeling*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1995.

[5] T. L. Chap, *Introductory Biostatistics*, John Wiley & Sons, Hoboken, NJ, USA, 2003.

[6] B. Vinzamuri, Y. Li, and C. K. Reddy, "Active Learning based survival regression for censored data," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM'14)*, pp. 241–250, Shanghai, China, November 2014.

[7] Y. Xiao and M. Watson, "Guidance on conducting a systematic literature review," *Journal of Planning Education and Research*, vol. 39, no. 1, pp. 93–112, 2017.

[8] C. Okoli, "A guide to conducting a standalone systematic literature review," *Communications of the Association for Information Systems*, vol. 37, no. 43, pp. 879–910, 2015.

[9] C. Okoli and K. Schabran, "A guide to conducting a systematic literature review of information system research," *Sprout*, vol. 10, pp. 10–26, 2010.

[10] K. Lynn, "Difference between a systematic review and a literature review," 2013, https://doi.org/10.6084/m9.figshare.766364.

[11] K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes, "Five steps to conducting a systematic review," *Journal of the Royal Society of Medicine*, vol. 96, no. 3, pp. 118–121, 2003.

[12] J. P. Klein and M. L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data*, Springer-Verlag, Berlin, Germany, 1997.

[13] W. J. Shih, "Problems in dealing with missing data and informative censoring in clinical trials," *Trials*, vol. 3, p. 4, 2002.

[14] L. H. Suttner, *Statistical methods for censored and missing data in Survival Analysis*, PhD dissertation, Penn Library, Wolverhampton, UK, 2019.

[15] M. Tableman and J. S. Kim, *Survival Analysis Using S*, Chapman & Hall, London, UK, 2004.

[16] 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.

[17] J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York, NY, USA, 1982.

[18] K.-M. Leung, R. M. Elashoff, and A. A. Afifi, "Censoring issues in survival analysis," *Annual Review of Public Health*, vol. 18, no. 1, pp. 83–104, 1997.

[19] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New York, NY, USA, 1980.

[20] M. M. Ngari, S. Schmitz, and C. Maronga, "A systematic review of the quality of conduct and reporting of survival analyses of tuberculosis outcomes in Africa," *BMC Medical Research Methodology*, vol. 21, p. 89, 2021.

[21] S. Selvin, *Survival Analysis for Epidemiology and Medical Research: A Practical Guide*, Cambridge University Press, New York, NY, USA, 2008.

[22] D. G. Kleinbaum and M. Klein, *Survival Analysis a Self-Learning Test*, Springer Science Business Media, New York, NY, USA, 2nd edition, 2005.

[23] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B*, vol. 34, no. 2, pp. 187.75–202.75, 1972.

[24] P. Diggle, P. Heagerty, K. Y. Liang, and S. Zeger, *Analysis of Longitudinal Data*, Oxford University Press, New York, NY, USA, 2nd edition, 2002.

[25] D. W. Hosmer Jr. and S. Lemeshow, *Applied Survival Analysis, Regression Modeling of Time to Event Data*, John Wiley & Sons, New York, NY, USA, 1999.

[26] E. T. Lee, *Statistical Methods for Survival Data*, Wadsworth, Belmont, CA, USA, 1980.

[27] K. Namboodiri and C. M. Suchindran, *Life Tables and Their Applications*, Academic Press, Orlando, FL, USA, 1987.

[28] J. D. Teachman, "Analyzing social processes: life tables and proportional hazards models," *Social Science Research*, vol. 12, no. 3, pp. 263–301, 1983.

[29] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, *Statistical Models Based on Counting Processes*, Springer, Berlin, Germany, 1993.

[30] T. R. Fleming and D. P. Harrington, "Counting processes and survival analysis," *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*, John Wiley & Sons, New York, NY, USA, 1991.

[31] Y. Kim and J. Seok, "Network estimation for censored time-to-event data for multiple events based on multivariate survival analysis," *PLoS One*, vol. 15, no. 10, Article ID e0239760, 2020.

[32] F. D. Atem, R. A. Matsouaka, and V. E. Zimmern, "Cox regression model with randomly censored covariates," *Biometrical journal. Biometrische Zeitschrift*, vol. 61, no. 4, pp. 1020–1032, 2019.

[33] M. David, J. Vock, S. Wolfson et al., "Adapting machine learning techniques to censored time-to- event health record data: a general-purpose approach using inverse probability of censoring weighting," *Journal of Biomedical Informatics*, vol. 61, pp. 119–131, 2016.

[34] T. M. P. Tran, S. Abrams, M. Aerts, K. Maertens, and N. Hens, "Measuring association among censored antibody titer data," *Statistics in Medicine*, vol. 40, no. 16, pp. 3740–3761, 2021.

[35] R. A. Canales, A. M. Wilson, J. I. Pearce-Walker, M. P. Verhougstraete, and K. A. Reynolds, "Methods for handling left-censored data in quantitative microbial risk assessment," *Applied and Environmental Microbiology*, vol. 84, no. 20, pp. 203–218, 2018.

[36] P. Soret, M. Avalos, L. Wittkop, D. Commenges, and R. Thiébaut, "Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 159, 2018.

[37] H. Støvring and I. S Kristiansen, "Simple parametric survival analysis with anonymized register data: a cohort study with truncated and interval censored event and censoring times," *BMC Research Notes*, vol. 4, p. 308, 2011.

[38] M. W. Donoghoe and V. Gebski, "The importance of censoring in competing risks analysis of the subdistribution hazard," *BMC Medical Research Methodology*, vol. 17, no. 1, p. 52, 2017.

[39] A. Yousefi, D. D. Dougherty, E. N. Eskandar, A. S. Widge, and U. T. Eden, "Estimating dynamic signals from trial data with censored values," *Computers in Psychiatry*, vol. 1, pp. 58–81, 2017.

[40] H. Ahmadi, D. A. Granger, K. R. Hamilton, C. Blair, and J. L. Riis, "Censored data considerations and analytical approaches for salivary bioscience data," *Psych Neuroendocrinology*, vol. 129, Article ID 105274, 2021.

[41] T. Yu, L. Wu, and P. Gilbert, "New approaches for censored longitudinal data in joint modelling of longitudinal and survival data with application to HIV vaccine studies," *Lifetime Data Analysis*, vol. 25, no. 2, pp. 229–258, 2019.

[42] H. R. Oliveira, S. P. Miller, L. F. Brito, and F. S. Schenkel, "Impact of censored or penalized data in the genetic evaluation of two longevity indicator traits using random regression models in north American angus cattle," *Animals*, vol. 11, no. 3, p. 800, 2021.

[43] B. Zupan, J. Demšar, M. W. Kattan, J. R. Beck, and I. Bratko, "Machine learning for survival analysis: a case study on recurrence of prostate cancer," *Artificial Intelligence in Medicine*, vol. 20, no. 1, pp. 59–75, 2000.

[44] I. Štajduhar and B. Dalbelo-Basić, "Learning bayesian networks from survival data using weighting censored instances," *Journal of Biomedical Informatics*, vol. 43, no. 4, pp. 613–622, 2010.

[45] N. C. Brownstein, V. Bunn, L. M. Castro, and D. Sinha, "Bayesian analysis of survival data with missing censoring indicators," *Biometrics*, vol. 77, no. 1, pp. 305–315, 2021.

[46] G. Gómez, M. L. Calle, and R. Oller, "Frequentist and Bayesian approaches for interval-censored data," *Statistical Papers*, vol. 45, no. 2, pp. 139–173, 2004.

[47] Z. Jianguo Sun and J. Sun, "Interval censoring," *Statistical Methods in Medical Research*, vol. 19, no. 1, pp. 53–70, 2010.

[48] S. Ji, L. Peng, R. Li, and M. J. Lynn, "Analysis of dependently censored data based on quantile regression," *Statistica Sinica*, vol. 24, no. 3, pp. 1411–1432, 2014.

[49] S. Portnoy, "Censored regression quantiles," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 1001–1012, 2003.

[50] W. Zhang, K. Chaloner, M. K. Cowles, Y. Zhang, and J. T. Stapleton, "A bayesian analysis of doubly censored data using a hierarchical cox model," *Statistics in Medicine*, vol. 27, no. 4, pp. 529–542, 2008.

[51] U. Domthong, C. R. Parikh, P. L Kimmel, and V. M. Chinchilli, "Assessing the agreement of biomarker data in the presence of left-censoring," *BMC Nephrology*, vol. 15, no. 1, p. 144, 2014.

[52] C. B. Guure and S. Bosomprah, "Bayesian perspective on randon censored survival data," *International Scholarly Research Notices*, vol. 2014, Article ID 430357, 9 pages, 2014.

[53] R. C. Gupta, O. Akman, and S. Lvin, "A study of log-logistic model in survival analysis," *Biometrical Journal*, vol. 41, no. 4, pp. 431–443, 1999.

[54] L. Peng, "Quantile regression for survival data," *Annual Review of Statistics and Its Application*, vol. 8, no. 1, pp. 413–437, 2021.

[55] T. Huynh, H. Quick, G. Ramachandran et al., "A comparison of the β-substitution method and a bayesian method for analyzing left-censored data," *Annals of Occupational Hygiene*, vol. 60, no. 1, pp. 56–73, 2016.