Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/9/20240706

From foundations to frontiers: Navigating survival analysis in the era of big data and deep learning

Jiaming Zhang

Vanderbilt University, Nashville, 27235, the United States

Jiaming.zhang@vanderbilt.edu

Abstract. In the era of big data, survival analysis, a statistical method for analyzing the expected duration of time until one or more events happen, has gained significant importance, especially in medical and biological research. This paper primarily focuses on the comprehensive exploration and understanding of survival analysis modelling, from traditional to modern approaches, and identifies the existing challenges and future prospects of these models. We commence by discussing foundational models such as the Kaplan-Meier and Cox proportional hazards models, and then transition into the exploration of the more flexible Accelerated Failure Time model. Acknowledging the current challenges faced in survival analysis, such as dealing with high-dimensional data, lack of labelled data, and data quality and reliability, we further delve into the potential solutions provided by modern techniques like deep learning, transfer learning, and semi-supervised learning. Additionally, the paper highlights the issues of interpretability and transparency of complex models, offering an overview of interpretability methods such as LIME and SHAP. Despite certain limitations, our study offers a valuable reference for understanding the evolution of survival analysis and sparks further discussions about its future development, emphasizing the profound significance of survival analysis in the realm of statistical research.

Keywords: survival analysis, Kaplan-Meier model, Cox proportional hazards model, accelerated failure time model, deep learning.

1. Introduction

Survival analysis is a method that allows for the examination and interpretation of the duration of survival in organisms or humans, using data collected from experiments or surveys. This approach is crucial for investigating the connections and intensity of impact between survival time, outcomes, and a range of influencing factors.

This statistical method primarily used for studying the probability of a specific event occurring at a certain time point or within a certain time frame, and this robust statistical tool, widely employed in diverse scientific disciplines such as biomedical studies, engineering, social sciences, and more, is dedicated to the analysis of time-to-event data.

Firstly, survival analysis, originating in biostatistics, has evolved significantly over the years. While early models laid foundational insights, contemporary methods, fueled by big data and computational growth, now extend to diverse sectors like finance and engineering. This section traces its developmental journey from inception to present-day applications.

DOI: 10.54254/2753-8818/9/20240706

Secondly, survival analysis has a rich history, originating from biostatistics with a focus on studying patterns of human survival times, such as disease survival rates, and in the development trajectory segment, this passage explores the novel research and data models that have emerged in the field of survival analysis in recent years.

Thirdly, in the future perspectives section, this research forecasts potential paths of survival analysis's development and new technologies and ideas that might have an impact. This includes factors such as model complexity, data diversity, and growth in computational power. Admittedly, future progress will also face certain impediments, so in the technical barriers section, this paper will discuss in detail the major technical challenges currently within survival analysis, issues that may hinder further advancements and improvements. The passage aims also to delve into how these challenges impact existing research efforts and potential strategies to overcome them.

This paper aims to provide a comprehensive overview of survival analysis, delving into its history, development trajectory, future perspectives, and existing technical barriers. By thoroughly examining these themes, this paper hopes to provide readers with a holistic perspective of survival analysis and a deep understanding of its future development. It is believed that understanding and applying survival analysis can not only propel the scientific progress in related fields, but also enhance researchers' perception and comprehension of the world.

2. Basic information of survival analysis

2.1. Censored data

Censored data plays a pivotal role in survival analysis, a branch of statistics that predominantly focuses on the time until the occurrence of an event of interest such as patient survival or equipment failure. However, in real-world observations, researchers often do not have the exact timing of all such events. Some events may occur after the observation period has ended, some may have occurred before the observation began, and for others, experimental operators only know that they occurred between two points of observation. This introduces the concept of censoring.

The types of censored data and how they are handled significantly impact the accuracy of survival analysis outcomes. Censored data can be classified in two primary ways: by types based on research design (Type I, II, III) and by directions based on observation (left, right, and interval censoring) [1].

Type-based classification pertains to the design of the study and data collection process. Type I censoring involves all subjects starting and ending the study at the same time, regardless of the number of events that have occurred. Type II censoring starts with all subjects at the same time, but the study ends when a predetermined number of events have taken place. Type III censoring allows subjects to enter the study at different times, but with a fixed end time for the study.

The choice between these types of censoring will affect data collection and, consequently, the results of survival analysis.

Direction-based classification, in addition, deals with the knowledge about the timing of the event. Left censoring applies when an event occurs before the observation begins; in case researchers know that the event has happened, but the experiments are lack of the exact time. Right censoring applies when an event occurs after the observation ends; researchers know that the event had not happened by the time that the observation ended, but similarly, experimental operators do not know when it occurred afterward. Interval censoring applies when an event occurs during the observation period, but researchers do not know the exact time, only that it occurred between two points of the entire observation. The correct identification and handling of censored data in different directions is a key to obtaining accurate results in survival analysis.

In summary, understanding and properly handling censored data is an essential component of survival analysis. Only by correctly dealing with censored data can researchers and experimenters achieve accurate and reliable results from survival analyses, which bears significant implications for clinical decision-making and policy development.

DOI: 10.54254/2753-8818/9/20240706

2.2. Survival time

Survival Time is the primary focus of study, defined as the time from the initiation of the study (or from a particular starting point like diagnosis or start of treatment) to the occurrence of the event of interest. This time-to-event data typically encompass two scenarios: the event has already occurred, and researchers record an exact time, or the event has not occurred by the end of the study, and researchers only have an upper limit of the event's occurrence, the latter being the previously discussed censored data [2].

2.3. Survival functions

The Survival Function (or survival probability) describes the probability of surviving, i.e., not experiencing the event of interest, at or before a given time point. The Survival Function is non-increasing, starting at 1 and gradually decreasing over time, signifying the increasing probability of the event occurring with the passage of time.

2.4. Hazard functions

The Hazard Function (or failure rate or instantaneous death rate) denotes the instantaneous probability of the event occurring at a given time. Unlike the Survival Function, the Hazard Function does not have a fixed range and can increase or decrease over time. The Hazard Function provides information about how the risk of the event happening varies over time.

2.5. Covariates

In survival analysis, a primary goal with covariates is to help people understand what factors may influence the survival time and to quantify their impact. This requires us to establish a model linking the survival time with one or more covariates, enabling us to explain and predict survival time. In this model, the coefficients of the covariates represent the relative change in survival time for a one-unit increase in the covariate. Positive coefficients imply that an increase in the covariate will increase survival time, whereas negative ones indicate that an increase in the covariate will reduce survival time. For instance, in clinical trials, certain research may be interested in covariates such as age, gender, type of treatment, etc. These covariates could influence the survival time of patients. By incorporating these covariates into the survival model, researchers can better understand how these factors affect survival time and conduct the experiments or research with a higher magnitude of their influence.

3. Early and modern survival analysis development and employments

Kaplan-Meier non-parametric estimation and semi-parametric estimation of Cox Proportional Hazards model proposed by D. R. Cox are the most widely used classical methods [3].

3.1. Kaplan-Meier non-parametric estimation

The Kaplan-Meier estimator is a non-parametric statistical method designed in 1958, primarily used for estimating the survival function from a specific starting time to the occurrence of a certain event. This method finds broad application in clinical trials and disease survival analysis. The Kaplan-Meier method can handle what is known as censored data, that is, situations where an event does not occur during the study period or subjects drop out of the study. This ability to handle censored data gives it an advantage in dealing with data that has varying times of censoring and event occurrence. Censored observations are subjects who either die of causes other than the disease of interest or are lost to follow-up [4].

As mentioned in the paper, if the estimation procedure allows for the best-fitting distribution to be chosen from all possible distributions, rather than restricting the choice to a specific category of distributions, then it would be reasonable to call this estimation procedure nonparametric [5]. Kaplan-Meier estimator's characteristics of non-parametric are reflected in the foundation of this method that it does not require the assumption that survival times follow a specific probability distribution.

The computation of the Kaplan-Meier estimator involves successively multiplying the survival probabilities across each time interval. The survival probability for each specific interval is ascertained

DOI: 10.54254/2753-8818/9/20240706

as the quotient of the count of individuals who endured within that interval over the total count of individuals under observation at the onset of that interval. Consequently, this approach facilitates robust estimation of the aggregate survival function, even amidst the occurrence of data censoring [5].

Considering the drawback of the Kaplan-Meier estimator, it is incapable of handling multiple risk factors inherently, necessitating other methods when considering various influences on survival time such as age, gender, or disease stage. Being a non-parametric method, it requires a substantial sample size for precise estimations. With smaller samples, the survival function estimation may exhibit significant fluctuations. Furthermore, the selection of time partitioning, crucial in the calculation of the estimator, could impact the shape of the survival function, especially when the occurrence of events in the data is sparse [5].

While the Kaplan-Meier estimator provides an essential foundation for survival analysis, its limitations suggest the necessity for more comprehensive methods when dealing with multifactorial influences on survival time, or when operating with smaller or more complex datasets. One approach that addresses these concerns and has found widespread use in survival analysis is the Cox Proportional Hazards Model. This method permits the simultaneous analysis of the impact of several variables on survival time and is not restricted by the same assumptions required by the Kaplan-Meier estimator. Therefore, it is necessary to delve deeper into the nuances of the Cox Proportional Hazards Model and explore how it complements the Kaplan-Meier estimator in survival analysis.

3.2. Cox proportional hazards model

The Cox Proportional Hazards Model was proposed by British statistician Sir David Cox in 1972 and has since become one of the most commonly used models in survival analysis. This method is used to evaluate the impact of multiple covariates on survival or event occurrence time. It has been widely applied in fields like clinical trials, epidemiological research, and economics for survival data analysis.

At the core of the Cox model is the proportional hazards assumption, which states that the covariates' effects on survival risk are multiplicative and remain constant over the observation period. This means that the hazard functions' ratio for any two individuals is a constant that does not depend on time. If this assumption does not hold, the model results may be biased. Therefore, it is crucial to test the proportional hazards assumption before employing the Cox model. The fundamental form of the Cox model is a semi-parametric model:

The first component, $h0(t)$, is the baseline hazard function. It represents the risk of the event happening at time t for an individual with all covariates equal to zero. It does not rely on any specific parameters and is not typically estimated in the Cox model, which instead focuses on the relative differences between individuals with different covariate values.

The second component, $\exp(X\beta)$, is the exponential of the linear predictor, and it adjusts the baseline hazard for the individual covariates. In this component, X represents the covariate values for an individual, andβrepresents the coefficients for those covariates (The coefficients of the covariates need to be estimated from the data using the method of Maximum Likelihood Estimation). Each covariate has its own coefficient, which measures the effect of that covariate on the hazard rate. If the coefficient is positive, then an increase in that covariate leads to an increase in the hazard rate (and thus a decrease in survival time). Conversely, if the coefficient is negative, an increase in the covariate leads to a decrease in the hazard rate (and thus an increase in survival time).

One key feature of the Cox model is that it does not make any assumptions about the form of $h0(t)$, which allows the model to be quite flexible. This is why it is referred to as a semi-parametric model: the baseline hazard function $h0(t)$ is non-parametric, and the covariate effects are parametric [6, 7].

The Cox model is broadly applicable across various scenarios. For instance, in tumor research, scientists might be interested in factors such as age, gender, tumor size, and treatment methods affecting patient survival time. In economics, investors might research how covariates such as education and economic status impact the duration of unemployment. The Cox model can assist in quantitatively describing these factors' influence on survival time [7].

DOI: 10.54254/2753-8818/9/20240706

The main advantages of the Cox model are that it can consider the effects of multiple covariates on survival time simultaneously and it does not require specific assumptions about the survival time distribution. Another advantage of the Cox model is that it can handle censored data, i.e., situations where the observed event time is truncated or partial information is missing. This is crucial in survival analysis as in many studies, during the experiments, researchers might not observe all individuals' event occurrence times. Some individuals might leave the study before its conclusion, or they might not have experienced the event by the end of the study. The Cox model can correctly handle such data, making it widely used in survival analysis.

However, the proportional hazards assumption is a significant limitation of the Cox model. If this assumption does not hold, the model's results may be biased. Additionally, the Cox model does not handle time-dependent covariates very well. Although the Cox survival analysis model has certain limitations and many other survival analysis methods have been proposed since, the Cox model remains one of the most commonly used and popular methods [8].

3.3. Accelerated failure time (AFT) model

Kaplan-Meier estimator and Cox proportional hazards model, undeniably, have made significant contributions to the field of survival analysis. They brought forth a fresh perspective to assess survival data, with the Kaplan-Meier estimator providing an empirical method to chart the survival curve and the Cox model introducing a semi-parametric approach that allows incorporating covariates without making strong assumptions about the form of baseline hazard.

However, an inherent limitation with these approaches lies in their inability to explicitly model the survival times, and the dependence of survival times on covariates is often non-intuitive and complex to interpret. The Cox model's dependence on the proportional hazard assumption, while allowing the introduction of covariates, might prove restrictive in certain scenarios, particularly when the hazards are not proportional [8].

Addressing these concerns, the AFT model offers a compelling alternative. The AFT model, in contrast to the Cox model, provides a more direct interpretation by assuming a parametric form for the survival distribution and modeling the effect of covariates multiplicatively on the survival time, hence intuitively expressing how covariates can "accelerate" or "decelerate" the occurrence of the event of interest. This directness, coupled with the model's flexibility in handling different underlying survival distributions and its ability to accommodate non-proportional hazards, makes the AFT model a versatile tool in modern survival analysis.

The AFT model is a commonly used statistical model in survival analysis, which directly models survival time. The primary assumption of the AFT model is that the influence of covariates on survival time can be depicted as "accelerating" or "decelerating" the "speed" of survival time, hence the name "Accelerated Failure Time Model".

In the AFT model, survival time is assumed to be composed of a deterministic part (determined by covariates) and a random part (determined by the error term).

$$\mathrm{T} = \mathrm{Y} * \exp(X\beta) \tag{1}$$

Here, T is the observed survival time, Y is the random survival time (following a certain known probability distribution), X is the matrix of covariates, and β represents the effect size of covariates.

A key feature of this model is its assumption that the effect of covariates on survival time is multiplicative, not additive. This means that the impact of covariates on survival time is achieved by multiplying a coefficient (i.e., $\exp(X\beta)$), rather than by adding or subtracting a quantity. This contrasts with the assumption of another common survival analysis model, the Cox proportional hazards model, which posits that the effect of covariates is multiplicative, influencing the hazard function, not the survival time itself.

In many demonstrations, the specific formula of the model is $\log Ti = Wi\beta + \varepsilon i$. One advantage of this is that it makes the model's expression closer to the linear regression model, simplifying computations and interpretation. It transforms the nonlinear relationship of the model into a linear one,

DOI: 10.54254/2753-8818/9/20240706

enabling the regression coefficient to be directly interpreted as an effect on the logarithm of survival time. The specific derivation simplifies $\exp(Xi\beta + \varepsilon i)$ to $Xi\beta + \varepsilon i$ [8].

3.4. A comparison of the AFT model with the Kaplan-Meier and Cox models

The AFT model directly models survival time and can intuitively explain the impact of covariates on survival time. The Kaplan-Meier and Cox models, on the other hand, focus more on modelling survival functions and hazard functions.

Compared to the Cox model, one advantage of the AFT model is that it has weaker assumptions about the relationship between covariates and survival time. The Cox model requires the proportional hazards assumption (that is, the influence of covariates is constant throughout the study period), while the AFT model does not have this requirement [8].

In addition, compared to the Kaplan-Meier model, the AFT model can handle continuous and categorical covariates, while the Kaplan-Meier model is mainly used to describe the overall distribution of survival time and cannot directly handle the impact of covariates.

The modern landscape of survival analysis has seen these models find broad applications, contributing to diverse areas such as biomedicine, engineering, and social sciences. More recently, novel techniques, such as machine learning and high-dimensional data handling, are being integrated into survival analysis, pushing its boundaries. The continuous development of these models, along with the integration of advanced analytical techniques, is paving the way for more sophisticated and nuanced understanding of time-to-event data, asserting the vibrant and dynamic future of survival analysis.

4. Technical barriers with some solutions

4.1. High-dimensional data

With the development of big data, analysis in this area often needs to deal with high-dimensional data, where traditional survival analysis methods may no longer be applicable. Therefore, mathematicians and other researchers need to develop new statistical methods and computational tools to handle high-dimensional data. In dealing with high-dimensional data, machine learning and deep learning algorithms have already shown great potential. For instance, deep learning methods can be used to learn higher-level abstract features of data and perform effective dimensionality reduction. Moreover, feature selection methods can also be applied to deal with high-dimensional data by filtering out the most important features to reduce data dimensions.

4.2. Lack of labelled data

In many circumstances, researchers may not have sufficient labelled data to perform survival analysis. For example, for rare diseases, they might not be able to collect enough case data. When facing the issue of lack of labelled data, transfer learning and semi-supervised learning are two possible solutions. Transfer learning is the application of knowledge learned in one domain (the source domain) to another domain (the target domain). Semi-supervised learning, on the other hand, involves using a small amount of labelled data and a large amount of unlabelled data for learning [9, 10].

4.3. Data quality and reliability

Due to irregularities in data collection, there might exist some erroneous, inaccurate, or missing data, all of which could impact the results of survival analysis. Adopting appropriate strategies during the data pre-processing stage can effectively handle issues related to data quality. For example, methods like data cleaning, anomaly detection, and data interpolation can be used to deal with erroneous, inaccurate, or missing data.

In survival analysis, complex machine learning models, such as deep learning and random forests, often outperform traditional statistical models in terms of prediction accuracy. However, the complexity of these models often reduces their interpretability. This is because their predictions are often based on intricate, hard-to-understand internal computations, making it difficult to comprehend how the model

DOI: 10.54254/2753-8818/9/20240706

operates and the basis of its predictions. In the medical, financial, and other fields that require explicit interpretation of model predictions, this lack of transparency and interpretability could be problematic.

To address this issue, researchers have proposed a range of methods to enhance the interpretability of machine learning models. These methods can be roughly divided into two categories: intrinsic interpretability methods and post-hoc interpretability methods.

Intrinsic interpretability methods primarily consider interpretability when designing the model. For example, decision trees and linear regression are naturally interpretable models. Another approach is to constrain model complexity to enhance interpretability, such as applying regularization terms to limit model complexity. Post-hoc interpretability methods, in addition interpret the workings of the model by analyzing its prediction results after training. Among them, Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive explanations (SHAP) are widely used post-hoc interpretation methods. LIME interprets the original model's predictions by generating interpretable models locally, while SHAP explains the model's predictions by calculating the contribution value of features [11].

In summary, although complex machine learning models face challenges in interpretability, satisfactory interpretive results can still be achieved through appropriate model design and post-hoc interpretation methods. Furthermore, this remains an active area of research, with the potential for more innovative methods to improve the interpretability of machine learning models in the future.

5. Conclusion

This comprehensive review traverses the journey of survival analysis, from the pioneering Kaplan-Meier and Cox models to contemporary Accelerated Failure Time (AFT) models. While traditional models have been instrumental in medical and statistical research, delineating the distribution of survival times and analyzing relationships between covariates and survival, they sometimes fall short due to inherent assumptions, such as proportional hazards. The AFT model, with its more lenient assumptions, offers a promising alternative in certain scenarios.

In the era of big data, survival analysis confronts new challenges: grappling with high-dimensional data, ensuring data reliability, and handling the scarcity of labelled data. The silver lining in this evolving landscape is the advent of innovative solutions like deep learning, transfer learning, and data pre-processing techniques that promise to mitigate these challenges. However, as the complexity of models heightens, ensuring their transparency and interpretability emerges as a crucial hurdle. This has led to the development of techniques like LIME and SHAP, specifically designed to enhance the clarity of intricate models.

Despite the thoroughness of this research, it isn't without limitations. The discourse leans heavily on statistical theories, side-lining practical challenges. Furthermore, while novel techniques are introduced, the paper lacks empirical studies validating their efficacy, which might restrict its real-world applicability. Nevertheless, this paper positions itself as a pivotal resource for understanding survival analysis, its evolution, and prospective trajectory. It aspires to catalyze further discussions and innovations in this domain.

References

[1] Emmert-Streib F., Dehmer M. Introduction to Survival Analysis in Practice. Machine Learning and Knowledge Extraction. 2019 Sep 8;1(3):1013–38. http://dx.doi.org/10.3390/make1030058

[2] Flynn R. Survival analysis. Journal of Clinical Nursing. 2012;21(19pt20):2789–97. doi:10.1111/j.1365-2702.2011.04023.

[3] Li Y., Zhao Q, Ma S. Recent Advances and Future Challenges for Biostatistics. Statistical Research. 2016 Jun;33. https://tjyj.stats.gov.cn/CN/10.19343/j.cnki.11-1302/c.2016.06.001

[4] D. Cameron Watt, Aitchison T, MacKie Rm, Sirel JM. Survival analysis: the importance of censored observations. 1996 Oct 1;6(5):379–85.

DOI: 10.54254/2753-8818/9/20240706

[5] Kaplan E.L., Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958 Jun;53(282):457–81. https://www.jstor.org/stable/2281868

[6] Cox D.R. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187–220. Available from: https://www.jstor.org/stable/2985181

[7] Deo S.V., Deo V., Sundaram V. Survival analysis—part 2: Cox proportional hazards model. Indian Journal of Thoracic and Cardiovascular Surgery. 2021 Jan 2; 37:229–33.

[8] Bradburn M.J., Clark T.G., Love S.B., Altman D.G. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. British Journal of Cancer, 2003 Aug 1;89(3):431–6. Available from: https://www.nature.com/articles/6601119#Sec3

[9] Bozinovski S. Reminder of the First Paper on Transfer Learning in Neural Networks, 19-76. Informatica. 2020 Sep 15;44(3). https://informatica.si/index.php/informatica/article/view/2828.

[10] Liu X., Zachariah D., Wågberg J., Schön T.B. Reliable Semi-Supervised Learning when Labels are Missing at Random. arXiv. 2019 Oct 24.

[11] Christoph Molnar. Interpretable machine learning: a guide for making Black Box Models interpretable. Morisville, North Carolina: Lulu, 2019.