Spring 2024

# Applications of Survival Estimation Under Stochastic Order to Cancer: The Three Sample Problem

Sage Vantine
svantine@mail.stmarytx.edu

## Recommended Citation

Vantine, S.A. (2024). Applications of Survival Estimation Under Stochastic Order to Cancer: The Three Sample Problem [Honors Thesis, St. Mary's University]. Digital Commons at St. Mary's University.

# Applications of Survival Estimation Under Stochastic Order to Cancer: The Three Sample Problem

by

Sage Vantine

HONORS THESIS

Presented in Partial Fulfillment of the Requirements for
Graduation from the Honors Program of
St. Mary's University
San Antonio, Texas

Dec 2023

Approved by

Kaitlin Hill

Dr. Kaitlin Hill
Department of Mathematics

Lori Boies

Dr. Lori Boies
Honors Program

# Applications of Survival Estimation Under Stochastic Order to Cancer: The Three Sample Problem

Sage Vantine

### Abstract

Stochastic ordering of probability distributions holds various practical applications. However, in real-world scenarios, the empirical survival functions extracted from actual data often fail to meet the requirements of stochastic ordering. Consequently, we must devise methods to estimate these distribution curves in order to satisfy the constraint. In practical applications, such as the investigation of the time of death or the progression of diseases like cancer, we frequently observe that patients with one condition are expected to exhibit a higher likelihood of survival at all time points compared to those with a different condition. Nevertheless, when we attempt to fit a survival curve based on real-world data, this anticipated behavior may not always be seen. Therefore, it becomes crucial to estimate these curves to provide a more accurate representation of the true survival times for patients under different conditions. To address this challenge, we harness the insights of various statisticians and adapt their methodologies from one-sample and two-sample cases to the more complex scenario of a three-sample case. In this case, we work with data obtained from three distinct populations for several kinds of cancer. Given the inherent complexity of such data, it is highly likely that the empirical survival functions derived from it will not conform to stochastic ordering constraints, necessitating the estimation process. This study investigates four different estimators applied to data representing the relative survival rates of various racial groups affected by eight different types of cancer. Ultimately, our goal is to determine which, if any, of these estimators perform the best in terms of Bias and Mean Square Error.

# Contents

## Acknowledgments

Special thanks to Dr. Kaitlin Hill and St. Mary's University. Preliminary research efforts were performed in partnership with Vanessa Avilez from California State University Fullerton and Evan Strong from Colorado Mesa University. I also want to thank my family, my roommates, my best friend Keily, and my dogs. Funding support for the preliminary research was provided by NSA grant number H98230-23-1-0017 and NSF grant DMS-2244093 to Dr. Javier Rojo, facilitated through the RUSIS@IU program.

## 1 Introduction

In real-world applications, the empirical survival functions extracted from actual data often defy the strict constraints of stochastic ordering. This disparity raises the necessity of finding methods to estimate the curves of these distributions, ensuring that the constraints of stochastic ordering are met. These estimations are instrumental in portraying the genuine survival times of distinct populations subject to different conditions, as exemplified in research focused on topics such as the time of death or the progression of diseases like cancer.

The challenges posed by real-world data and the vital role of stochastic ordering converge in this study, where we look at the insights of various statisticians. The objective is to adapt and expand their methodologies, originally designed for simpler one-sample and two-sample cases, into the more intricate three-sample scenario. We do this by applying these methods to data collected from three unique populations. Given the inherent complexity of this data, it is foreseeable that the empirical survival functions derived may deviate from the stipulated constraints, underscoring the need for innovative estimation techniques.

This research aims to investigate four distinct estimators, which are put to the test against real-world data representing the relative survival rates of diverse racial groups afflicted by eight different forms of cancer. The aim is to scrutinize these estimators, employing metrics like Bias and Mean Squared Error, to decide which, if any, stands as the optimal choice for future survival data analysis.

Dr. Javier Rojo addressed the challenge of estimating two empirical survival curves derived from samples taken from two stochastically ordered distributions. He utilized estimators proposed by Shaw-Hwa Lo (Lo, 1987; Rojo, 2004). Extending these concepts to the three-sample scenario poses a unique difficulty, as there is no universally recognized "best" method for estimating empirical survival functions under this scenario. While El Barmi and Mukerjee explored cases with $k > 2$ populations in 2005, they did not provide a specific estimation approach tailored to exactly three samples (Barmi & Mukerjee, 2005).

This paper discusses estimators generated through collaboration with Rojo and one proposed by El Barmi and Mukerjee's k-sample estimator (Barmi & Mukerjee, 2005). Other methods have been explored to estimate these empirical survival functions to maintain stochastic ordering, such as non-parametric maximum likelihood estimation by Park, Kalbfleisch, and Taylor, and empirical likelihood tests by El Barmi (El Barmi, 2017; Park, Kalbfleisch, & Taylor, 2012).

In this research, we employ diverse techniques to select the maximum and minimum values from one, two, or all three empirical survival functions, thereby effectively addressing the stochastic ordering constraint. Leveraging empirical survival functions is particularly beneficial for estimating true survival functions due to their asymptotic distribution and unbiased nature. To assess the performance of the four estimators applied in this study, we evaluate their bias and mean squared error (MSE).

In essence, this paper seeks to identify the most effective estimator among four options for estimating survival curves in the context of three distinct populations: non-Hispanic Black, non-Hispanic White, and Hispanic individuals—across various types of cancer. The sections are structured as follows: this first section acts as an introduction, the second section delves into essential background information, the third section describes the chosen estimators, the fourth section offers a concise overview of preliminary research conducted during the Summer of 2023 at the RU-SIS@IU program, the fifth section details the project's methodology, and the sixth section presents a thorough analysis of the results.

# 2 Background

The following discussion provides a comprehensive background, introducing the core principles of survival analysis and the concept of stochastic ordering.

We begin by introducing the concept of survival analysis and its application in various domains. From there, we delve into stochastic ordering, an important concept in survival analysis, and explain its significance in comparing and ordering probability distributions. While stochastic ordering provides a robust framework for theoretical comparisons, the empirical reality often presents unique challenges. Real-world data may deviate from what is expected, and these discrepancies require innovative methods for accurate estimation of survival curves.

Accurate survival curve estimation has the potential to transform healthcare, epidemiology, and statistical sciences, contributing to more informed decisions and improved models. Our investigation addresses these challenges, aiming to select the most suitable estimator for future survival data analyses while recognizing the critical role of statistical science in addressing real-world complexities.

## 2.1 Introduction to Survival Analysis

Survival analysis is a statistical methodology that plays a pivotal role in understanding and modeling the time to an event of interest, particularly in contexts where time is a fundamental aspect of the study (Oakes, 2000). Its core objective is to investigate the distribution of time until a specific event occurs, which can encompass a variety of outcomes, from life and death in medical studies to failure times of mechanical components in engineering.

In healthcare, for example, survival analysis is frequently used to explore the progression of diseases, estimate patient lifetimes, and compare the efficacy of different treatments. In actuarial science, it is essential for determining life insurance premiums and the associated risk assessments. The survival analysis framework offers a unique perspective on the uncertainties inherent in various events and outcomes, making it essential in addressing research questions related to the duration of time-based processes.

## 2.2 Stochastic Ordering

A central concept in survival analysis is stochastic ordering. Stochastic ordering entails a comparison between probability distributions, offering a framework for understanding the relative risks and lifetimes of different populations or groups. It enables researchers to quantify the probability that one event occurs earlier than another, given the distributions of these events.

Stochastic ordering establishes a hierarchy among random variables by characterizing their relative "favorability." Specifically, given two probability distributions, one can be consistently more "favorable" than the other. Stochastic ordering allows the systematic exploration of which event occurs sooner or more frequently.

We let $X$ and $Y$ represent two random lifetimes and let $X$ and $Y$ have $F$ and $G$ as their distinct distribution functions, respectively. E.L. Lehmann defines $X$ as being stochastically larger than $Y$ if $F(x) \leq G(x)$ for all $x$. In terms of survival functions, $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$, stochastic order must be estimated subject to the condition that $\bar{F}(x) \geq \bar{G}(x)$ (Rojo, 2004).

## 2.3 Real-World Challenges

While stochastic ordering serves as a powerful theoretical construct in survival analysis, the real-world application of these principles often presents challenges. One of the primary challenges arises from the discrepancies observed between theoretical expectations and empirical observations. Empirical survival functions derived from actual data may not always conform to the constraints imposed by stochastic ordering. As mentioned before, this emphasizes the need for methodologies to estimate the true survival curves of diverse populations under different conditions accurately.

These challenges are particularly pronounced in practical applications, such as studying the time of death or the progression of diseases like cancer. In these scenarios, it is frequently observed that patients with certain conditions are expected to exhibit higher likelihoods of survival at all time points compared to those with different conditions. However, when attempting to fit a survival curve based on real-world data, this anticipated behavior may not always be realized.

## 2.4 Literature Review

The estimation of distribution functions subject to stochastic order constraints is a critical problem in statistical research. This literature review synthesizes the findings and methodologies presented in a selection of scholarly articles that address this issue. Stochastic order constraints, which dictate the relationship between two or more distribution functions, have applications in various fields, such as reliability analysis, finance, and epidemiology. This review highlights the key contributions from each article, emphasizing their relevance to statistical estimation and their implications for practical applications.

Arcones et al. introduced the concept of stochastic precedence between random variables $X$ and $Y$ and discussed estimation methods when data are subject to such constraints (Arcones et al., 2002). Although their definition of stochastic precedence differs from other studies, it remains comparable. The authors proposed two estimation approaches: data shrinkage and data translation. These methods yield estimators that adhere to the stochastic precedence constraint and are shown to be root $n$-consistent (Arcones et al., 2002). This work provides essential insights into estimation methods for data with stochastic precedence constraints, which are valuable for understanding the behavior of random variables in various contexts.

In a journal article written in 2017, El Barmi focused on testing for uniform stochastic ordering between two univariate distribution functions using empirical likelihood-based tests, specifically under right censoring, which indicates that the survival times for certain individuals extend beyond the observation period or certain individuals died from events other than that being studied, introducing complexities that need to be addressed in the analysis (El Barmi, 2017). These tests, based on localized empirical likelihood statistics, exhibit distribution-free asymptotic properties (El Barmi, 2017). The study demonstrated the superiority of these tests in terms of power compared to traditional methods. This research contributes valuable tools for assessing stochastic ordering in practical scenarios with censored data.

El Barmi and Mukergee extended the study of stochastic ordering to cases involving multiple populations, each characterized by a distribution function (Barmi & Mukerjee, 2005). They provided simpler estimators for situations with more than two populations ($k > 2$) and demonstrated strong and uniform consistency (Barmi & Mukerjee, 2005). This work is particularly relevant for researchers studying scenarios with multiple populations under stochastic ordering constraints.

Lo addressed the estimation of two sets of data, each following independent distribution functions, while maintaining an order restriction that one distribution is stochastically greater than the other (Lo, 1987). The paper established asymptotic minimax bounds and constructed estimators that adhere to the specified order relationship (Lo, 1987). This approach is applicable to scenarios where maintaining stochastic order constraints is essential for accurate distribution function estimation.

Park et al. proposed a pointwise-constrained non-parametric maximum likelihood estimator for survival functions under right censoring and stochastic order constraints (Park, Taylor, & Kalbfleisch, 2012). This novel estimator outperformed alternative methods in both small and large sample scenarios, as demonstrated through simulations and the analysis of real-world data (Park, Taylor, & Kalbfleisch, 2012). The approach offers a practical solution for estimating survival functions while preserving stochastic order relationships.

Puri and Singh presented an estimator for an unknown cumulative distribution function (CDF) when a known CDF stochastically dominates the estimator (Puri & Singh, 1992). This estimator, based on modifying the empirical CDF, was proven to be consistent and demonstrated smaller mean squared error compared to the standard empirical CDF (Puri & Singh, 1992). This approach aligns with several other studies in this review, emphasizing the importance of maintaining order constraints in distribution function estimation.

Rojo and El Barmi introduced a family of estimators for survival functions under second-order stochastic dominance (Jiménez & Barmi, 2003). These estimators demonstrated strong uniform consistency and outperformed the empirical distribution function for specific loss functions (Jiménez & Barmi, 2003). The study expanded the understanding of distribution function estimation under second-order stochastic dominance, offering practical applications in fields where such dominance relationships are relevant.

Rojo and Ma explored nonparametric maximum likelihood estimators in the context of estimating survival functions under stochastic ordering constraints (Rojo & Ma, 1996). The study compared these estimators with others proposed in previous research, highlighting potential biases associated with different methods, particularly in one-sample problems (Rojo & Ma, 1996). This research informs the choice of estimation methods when dealing with stochastic ordering constraints in survival analysis.

Rojo discussed the estimation of distribution functions under stochastic order constraints, introducing new estimators that are strongly uniformly consistent (Rojo, 2004). These estimators are particularly useful when one of the sample sizes increases. The paper also examined the asymptotic distribution of these estimators for hypothesis testing purposes (Rojo, 2004). The findings emphasized the superior performance of these estimators in various examples, underscoring their practical significance.

# 3 The Estimators

This section will describe, in detail, the four estimators that were used in this study. The first three estimators were proposed by Dr. Javier Rojo from Indiana University during the 2023 RUSIS@IU program, while the fourth estimator was chosen in collaboration with Dr. Rojo during the same program. The goal of the estimators was to use varying techniques to estimate stochastic order. These techniques include using the max and min functions, as well as benchmark functions that will be discussed later in this section. Also used in this research are empirical survival functions, which we define as the estimate of a true, but often unknown, survival function based on observed data.

Each estimator assumes the existence of three populations from which we have three empirical survival functions $\bar{F}^*$, $\bar{G}^*$, and $\bar{H}^*$ where at most points $\bar{F}^* \geq \bar{G}^* \geq \bar{H}^*$, in other words the conditions for stochastic order are nearly met.

We define the benchmark function $R$ as the weighted average between two empirical survival functions as follows:

$$R_{n_1n_2}(\bar{F}^*, \bar{G}^*) = \frac{n_1 \cdot \bar{F}^*(x) + n_2 \cdot \bar{G}^*(x)}{n_1 + n_2} \tag{1}$$

$$R_{n_2n_3}(\bar{G}^*, \bar{H}^*) = \frac{n_2 \cdot \bar{G}^*(x) + n_3 \cdot \bar{H}^*(x)}{n_2 + n_3} \tag{2}$$

$$R_{n_1n_3}(\bar{F}^*, \bar{H}^*) = \frac{n_1 \cdot \bar{F}^*(x) + n_3 \cdot \bar{H}^*(x)}{n_1 + n_3} \tag{3}$$

The estimation of each survival function will be denoted as $\hat{\bar{F}}$, $\hat{\bar{G}}$, and $\hat{\bar{H}}$.

## 3.1 First Estimator

The first estimator was suggested by Dr. Javier Rojo during the 2023 RUSIS@IU program. This estimator depends on the maximum and minimum of the empirical survival probabilities. Estimator 1 is a mathematical method designed to estimate three survival functions, denoted as $\hat{\bar{F}}$, $\hat{\bar{G}}$, and $\hat{\bar{H}}$, based on empirical survival functions $\bar{F}^*$, $\bar{G}^*$, and $\bar{H}^*$ derived from three populations. The key assumption is that the stochastic order is nearly met, meaning that at most points, the survival probabilities follow the order $\bar{F}^* \geq \bar{G}^* \geq \bar{H}^*$. The estimation involves a benchmark function $R$ defined as weighted averages between pairs of empirical survival functions. Estimator 1 for $\hat{\bar{F}}$ considers three scenarios based on the maximum survival probability among the populations at a given point, incorporating weighted averages and benchmark functions accordingly. Similarly, for $\hat{\bar{H}}$, three scenarios are considered based on the minimum survival probability. We define Estimator 1 as follows:

$$\hat{\bar{F}}(x) = \begin{cases} \bar{F}^*(x) & \text{if } \max\{\bar{F}^*(x), \bar{G}^*(x), \bar{H}^*(x)\} = \bar{F}^*(x), \\ \max\{R_{n_1 n_2}(\bar{F}^*, \bar{G}^*), R_{n_1 n_3}(\bar{F}^*, \bar{H}^*)\} & \text{if } \max\{\bar{F}^*(x), \bar{G}^*(x), \bar{H}^*(x)\} = \bar{G}^*(x)), \\ R_{n_1 n_3}(\bar{F}^*, \bar{H}^*) & \text{if } \max\{\bar{F}^*(x), \bar{G}^*(x), \bar{H}^*(x)\} = \bar{H}^*(x), \end{cases} \tag{4}$$

$$\hat{\bar{H}}(x) = \begin{cases} \bar{H}^*(x) & \text{if } \min\{\bar{F}^*(x), \bar{G}^*(x), \bar{H}^*(x)\} = \bar{H}^*(x), \\ \min\{R_{n_1 n_3}(\bar{F}^*, \bar{H}^*), R_{n_2 n_3}(\bar{G}^*, \bar{H}^*)\} & \text{if } \min\{\bar{F}^*(x), \bar{G}^*(x), \bar{H}^*(x)\} = \bar{G}^*(x), \\ R_{n_1 n_2}(\bar{F}^*, \bar{H}^*) & \text{if } \min\{\bar{F}^*(x), \bar{G}^*(x), \bar{H}^*(x)\} = \bar{F}^*(x), \end{cases} \tag{5}$$

and

$$\hat{\bar{G}}(x) = \max\{\hat{\bar{H}}(x), \min\{\hat{\bar{F}}(x), \bar{G}^*(x)\}\}. \tag{6}$$

Notably, the use of the *max* function in $\hat{\bar{F}}$ and the *min* function in $\hat{\bar{H}}$, when $\bar{G}^*$ is the maximum or minimum, ensures the estimator adapts to the relative positions of the populations at specific points. The significance lies in the nuanced treatment of survival probabilities, leveraging benchmark functions and weighted averages to capture the interplay between the populations in the estimation process.

## 3.2 Second Estimator

The second estimator suggested by Dr. Rojo was inspired by the one-sample problem considered by Rojo and Ma, where the stochastically smallest empirical CDF is treated as a known distribution (Rojo & Ma, 1996). Estimator 2 is a straightforward and comparably simple method for estimating three survival functions. In this approach, each estimator is determined by selecting the maximum survival probability among the corresponding populations at a given point. Specifically, $\hat{\hat{F}}$ is the maximum among $\bar{F}^*$, $\bar{G}^*$, and $\bar{H}^*$, $\hat{\hat{G}}$ is the maximum between $\bar{G}^*$ and $\bar{H}^*$, and $\hat{\hat{H}}$ is simply the survival function for $\bar{H}^*$. Unlike Estimator 1, this method does not involve weighted averages. Instead, it provides a straightforward estimation by selecting the survival function associated with the population that exhibits the highest survival probability at each point. This simplicity may be useful in situations where a direct and unweighted comparison is favorable. We define the second estimator as follows:

$$\hat{\hat{F}}(x) = \max\{\bar{F}^*(x), \bar{G}^*(x), \bar{H}^*(x)\}, \tag{7}$$

$$\hat{\hat{G}}(x) = \max\{\bar{G}^*(x), \bar{H}^*(x)\} \tag{8}$$

$$\hat{\hat{H}}(x) = \bar{H}^*(x). \tag{9}$$

## 3.3 Third Estimator

The third estimator was also suggested by Dr. Javier Rojo during the 2023 RUSIS@IU program. The estimation process for this estimator involves a combination of direct selection and use of benchmark functions. Specifically, $\hat{\hat{F}}$ is determined by selecting the maximum survival probability between $\bar{F}^*$ and the estimated $\hat{\hat{G}}$. $\hat{\hat{G}}$ itself is determined by selecting the maximum between $\bar{G}^*$ and a benchmark function, $R_{n_2 n_3}(\bar{G}^*, \bar{H}^*)$. This benchmark function is a weighted average between the empirical survival functions of $\bar{G}^*$ and $\bar{H}^*$, introducing a comparative element in the estimation. Similarly, $\hat{\hat{H}}$ is determined by selecting the minimum survival probability between $\bar{H}^*$ and the same benchmark function. Estimator 3, therefore, provides a more nuanced approach to survival function estimation. We define Estimator 3 as follows:

$$\hat{\hat{F}}(x) = \max\{\bar{F}^*(x), \hat{\hat{G}}(x)\}, \tag{10}$$

$$\hat{\hat{G}}(x) = \max\{\bar{G}^*(x), R_{n_2 n_3}(\bar{G}^*, \bar{H}^*) \tag{11}$$

$$\hat{\hat{H}}(x) = \min\{\bar{H}^*(x), R_{n_2 n_3}(\bar{G}^*, \bar{H}^*)\}. \tag{12}$$

## 3.4 Fourth Estimator

The fourth estimator considers the benchmark of all three estimators, inspired by El Barmi and Mukerjee (Barmi & Mukerjee, 2005). The estimator they propose is defined by $\hat{\hat{F}}_i = \max_{r \leq i} \min_{s \geq i} \sum_{j=r}^{s} \frac{n_j \bar{F}_j}{\sum_{j=r}^{s} n_j}$ (Barmi & Mukerjee, 2005) where $1 \leq r, s \leq k$. Each estimator is defined differently based on the bounds of the maximum and minimum.

For the purposes of this research, we create a new benchmark function that includes more than two of the survival functions. The combined weighted average of all three empirical survival functions will be denoted as $R_{n_1 n_2 n_3}(\bar{F}^*, \bar{G}^*, \bar{H}^*)$ and defined as

$$R_{n_1 n_2 n_3}(\bar{F}^*, \bar{G}^*, \bar{H}^*) = \frac{n_1 \bar{F}^*(x) + n_2 \bar{G}^*(x) + n_3 \bar{H}^*(x)}{n_1 + n_2 + n_3}, \tag{13}$$

using the same earlier defined empirical survival functions.

For this estimator, the estimation process for each survival function is defined in terms of a selection among various benchmark functions. Specifically, $\hat{\hat{F}}$ is determined by selecting the maximum among the benchmark functions involving all possible pairs and the newly introduced triplet of survival functions. $\hat{\hat{G}}$ is directly set equal to the benchmark function representing the combined weighted average of all three populations. Lastly, $\hat{\hat{H}}$ is determined by selecting the minimum among the benchmark functions. Estimator 4 is notable for incorporating both two-way and three-way comparisons between the survival functions, offering a refined approach to survival function estimation in the context of multiple populations.

Estimator 4 will be defined as follows:

$$\hat{\hat{F}}(x) = \max\{R_{n_1 n_2}(\bar{F}^*, \bar{G}^*), R_{n_2 n_3}(\bar{G}^*, \bar{H}^*), R_{n_1 n_3}(\bar{F}^*, \bar{H}^*), R_{n_1 n_2 n_3}(\bar{F}^*, \bar{G}^*, \bar{H}^*)\}, \tag{14}$$

$$\hat{\hat{G}}(x) = R_{n_1 n_2 n_3}(\bar{F}^*, \bar{G}^*, \bar{H}^*), \tag{15}$$

$$\hat{\hat{H}}(x) = \min\{R_{n_1 n_2}(\bar{F}^*, \bar{G}^*), R_{n_2 n_3}(\bar{G}^*, \bar{H}^*), R_{n_1 n_3}(\bar{F}^*, \bar{H}^*), R_{n_1 n_2 n_3}(\bar{F}^*, \bar{G}^*, \bar{H}^*)\}. \tag{16}$$

# 4 Preliminary Research

Preliminary findings from collaborative research conducted by Vanessa Avilez of California State University Fullerton and Evan Strong of Colorado Mesa University, as part of the Research Experience for Undergraduates (REU) program RUSIS at Indiana University during the summer of 2023, have provided valuable insights into the theoretical performance of three-sample estimators in both censored and uncensored data scenarios. The generated data, drawn from diverse distributions, served as the testing ground for the four estimators described earlier.

In this research, we delved into the intricacies of a three-sample scenario, exploring both censored and uncensored data. We engaged the same estimators described above in rigorous comparisons, aiming to identify the one that performed best. Our observations led us to an intriguing conclusion: in the realm of uncensored data, no single estimator emerged as the unequivocal “champion." However, when dealing with right-censored data, a compelling victor did emerge, and it was none other than the simplest of the estimators, Estimator 2. This finding underscores the practical utility of straightforward methodologies in scenarios where data is subject to right censoring, a common reality in real-life datasets.

In the context of the uncensored case, Estimator 2 exhibited notable shortcomings in estimating $\hat{\hat{G}}$ and $\hat{\hat{H}}$, although it performed well for $\hat{\hat{F}}$. However, for right censored data, Estimator 2 emerged as the consistently superior choice across various simulated distributions. This finding carries practical significance, particularly when applied to real datasets where right censoring is commonplace. The simplicity of Estimator 2, involving the maximization of Kaplan-Meier functions and treating one of them as a known distribution, demonstrated robust performance in scenarios where data is subject to censoring. These insights lay the groundwork for a more nuanced understanding of estimator performance in real-world applications.

Given the prevalence of censoring in real-life datasets, the prospects for a simplified estimator to excel offer exciting possibilities for future applications and investigations.

In the research that is the focus of this study, it was expected that applying these estimators to real-life data will have similar results as real data is more often than not censored.

# 5 Methods

The data for this study was obtained from real-world cases involving three distinct populations: non-Hispanic Black, non-Hispanic White, and Hispanic individuals. The data pertains to the survival rates of various types of cancer within these populations. The data was collected from the National Cancer Institute's Surveillance, Epidemiology, and End Results Program (SEER) ("Surveillance, Epidemiology, and End Results (SEER) Program", 2023). The key variables of interest were the survival times for individuals within each population, with stratification by race and cancer type.

Empirical survival models were fitted to the data. These models are essential for understanding the baseline survival probabilities for each population and cancer type based on the raw data.

The four different estimators described above were then applied to approximate the survival functions for the three populations. Bias and Mean Squared Error (MSE) were then computed to assess the performance of each estimator. These measures were used to evaluate the extent to which each estimator deviates from the true survival probabilities given by the data.

Bias was calculated as the difference between the estimated survival probabilities and the true survival probabilities for each estimator, as follows:

$$\text{Bias} = \hat{\bar{F}} - \bar{F}^*. \tag{17}$$

The bias provides insight into the direction and magnitude of estimation errors.

The Mean Squared Error (MSE) was used to quantify the overall accuracy of the estimators by considering both bias and variance. It was calculated as the mean of the squared differences between the estimated and true survival probabilities, as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\bar{F}}(x_i) - \bar{F}^*(x_i))^2. \tag{18}$$

The purpose of using Mean Square Error is to quantify the accuracy of the estimators by considering bias and variance.

To visually represent the findings, various graphs and tables were generated. These visualizations include Empirical Survival Graphs displaying empirical survival probabilities for each population and cancer type, Bias Graphs illustrating the bias of each estimator over time, and a table summarizing the MSE values for each estimator for each cancer.

Based on the results of the bias and MSE calculations and the visualizations, conclusions were drawn regarding which estimator provided the best fit for the survival data within the studied populations.

# 6 Results

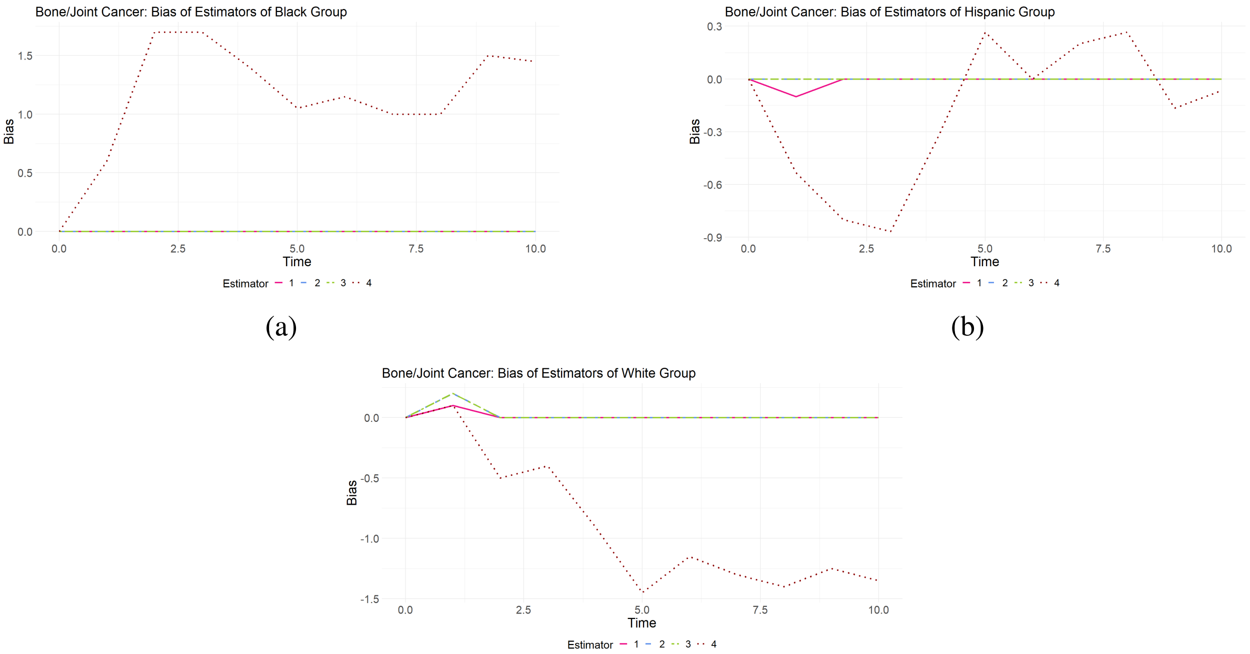
This section delves into an analysis of the performance of the four distinct estimators in this study. For a visualization of the original empirical survival data for each cancer type, see the Appendix. This section investigates the estimators' efficacy in approximating survival functions. The goal is to decide which estimator emerged as the optimal fit for the survival data in the studied populations.

## 6.1 Bones and Joint Cancer

| | Hispanic | Non-Hispanic Black | Non-Hispanic White |
|---|---|---|---|
| Estimator 1 | 0.00091 | 0.00000 | 0.00091 |
| Estimator 2 | 0.00000 | 0.00000 | 0.00364 |
| Estimator 3 | 0.00000 | 0.00000 | 0.00364 |
| Estimator 4 | 0.18192 | 1.53432 | 1.06273 |

Table 1: Table representing Mean Square Error (MSE) values for all estimators within the Bone/Joint Cancer category across the three populations: Non-Hispanic Black, Hispanic, and Non-Hispanic White.

We first studied survival rates for cancer of the bones and joints. The raw estimator results have been included in the appendices for each cancer type. For Bone and Joint Cancer estimator-based survival functions, see Figure 7 in the Appendix. The examination of Bone and Joint Cancer data reveals Estimator 4 as exhibiting the highest bias across all three populations, which is evident in Figure 1, indicating a notable magnitude of estimation errors. Concurrently, Estimator 4 records the highest MSE values, depicted in Table 1, demonstrating its suboptimal performance in all estimations for this cancer type. Remarkably, Estimator 2 and Estimator 3 share identical MSE values, as illustrated in Table 1, and exhibit parallel bias patterns across the three racial populations, as seen in Figure 1. Meanwhile, Estimator 1 showcases low bias and MSE values, making it challenging to definitively determine its superiority over Estimators 2 and 3 in this context. Estimator 4 emerges as the least accurate, exhibiting the highest bias and MSE values across all three populations. In contrast, Estimator 2 and Estimator 3, despite their simplicity, demonstrate comparable and superior performance, making it challenging to definitively declare one as most superior.



(a)

(b)

(c)

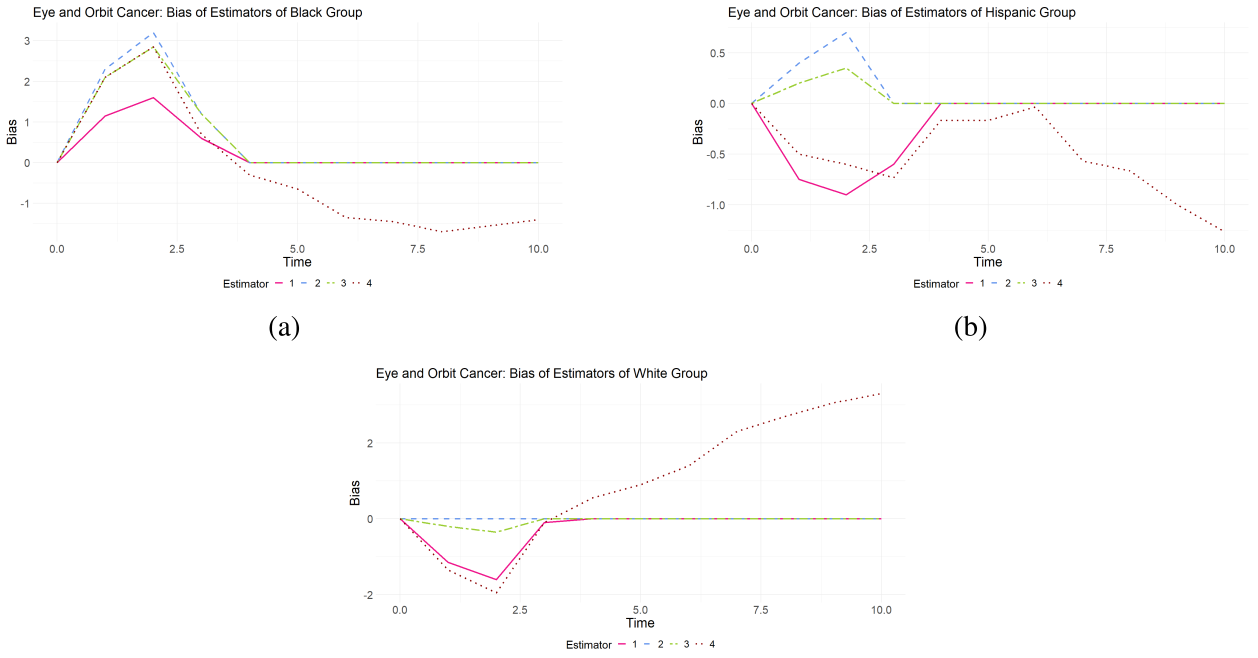
Figure 1: Visual representation of bias trends for all estimators within the Bone/Joint Cancer category across distinct populations, including (a) Non-Hispanic Black, (b) Hispanic, and (c) Non-Hispanic White groups.

## 6.2 Eye and Orbit Cancer

| | Hispanic | Non-Hispanic Black | Non-Hispanic White |
|---|---|---|---|
| Estimator 1 | 0.15750 | 0.38568 | 0.35386 |
| Estimator 2 | 0.05909 | 1.54273 | 0.00000 |
| Estimator 3 | 0.01477 | 1.27023 | 0.01477 |
| Estimator 4 | 0.41586 | 2.24659 | 3.77091 |

Table 2: Table representing Mean Square Error (MSE) values for all estimators within the Eye/Orbit Cancer category across the three populations: Non-Hispanic Black, Hispanic, and Non-Hispanic White.

The examination of data related to Eye and Orbit Cancer yielded intriguing findings. For Eye and Orbit Cancer estimator-based survival functions, see Figure 8 in the Appendix. Estimator 3 exhibited the lowest MSE value for the Hispanic Group, Estimator 1 achieved the lowest MSE for the Non-Hispanic Black Group, and Estimator 2 secured the lowest MSE for the Non-Hispanic White Group, as detailed in Table 1. Likewise, the bias analysis in Figure 2 highlights that Estimator 1 displayed the least bias for the Non-Hispanic Black Group, Estimator 3 exhibited the least bias for the Hispanic Group, and Estimator 2 demonstrated a bias of 0 for the Non-Hispanic White Group. Overall, determining the most superior performer among the estimators in this context proves challenging. In this case, each estimator exhibits strengths, with Estimators 1, 2, and 3 showcasing low bias and competitive MSE values across different racial groups. Determining a singular superior performer proves challenging.



(a)

(b)

(c)

Figure 2: Visual representation of bias trends for all estimators within the Eye/Orbit Cancer category across distinct populations, including (a) Non-Hispanic Black, (b) Hispanic, and (c) Non-Hispanic White groups.

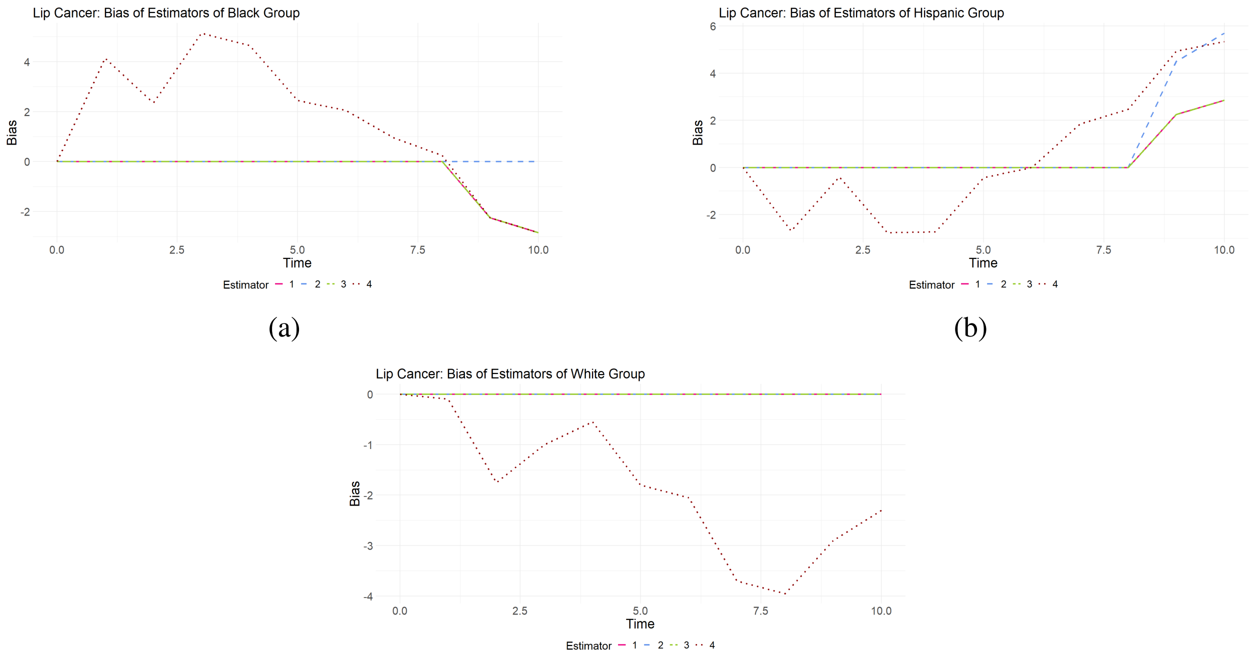
## 6.3 Lip Cancer

| | Hispanic | Non-Hispanic Black | Non-Hispanic White |
|---|---|---|---|
| Estimator 1 | 1.19864 | 1.19864 | 0.00000 |
| Estimator 2 | 4.79455 | 0.00000 | 0.00000 |
| Estimator 3 | 1.19864 | 1.19864 | 0.00000 |
| Estimator 4 | 7.72647 | 8.65864 | 4.98273 |

Table 3: Table representing Mean Square Error (MSE) values for all estimators within the Lip Cancer category across the three populations: Non-Hispanic Black, Hispanic, and Non-Hispanic White.

For Lip Cancer estimator-based survival functions, see Figure 9 in the Appendix. The examination of Lip Cancer data also consistently highlights Estimator 4 as the least accurate among the four estimators, evident in the pronounced bias depicted in Figure 3. Table 3 further reinforces this observation, indicating that Estimator 4 exhibits the highest MSE values across all three racial populations, affirming its inferior performance. Conversely, MSE values in Table 3 showcase a competitive scenario among Estimators 1, 2, and 3, with ties for the smallest MSE within specific racial groups. Notably, the first three estimators consistently exhibit minimal bias. Despite this result, pinpointing the superior performer remains challenging in this context.

Lip Cancer data consistently highlights Estimator 4 as the least accurate, with the highest bias and MSE values. Estimators 1, 2, and 3, on the other hand, present a competitive scenario with minimal bias, making it difficult to pinpoint a clear superior performer.



(a)

(b)

(c)

Figure 3: Visual representation of bias trends for all estimators within the Lip Cancer category across distinct populations, including (a) Non-Hispanic Black, (b) Hispanic, and (c) Non-Hispanic White groups.

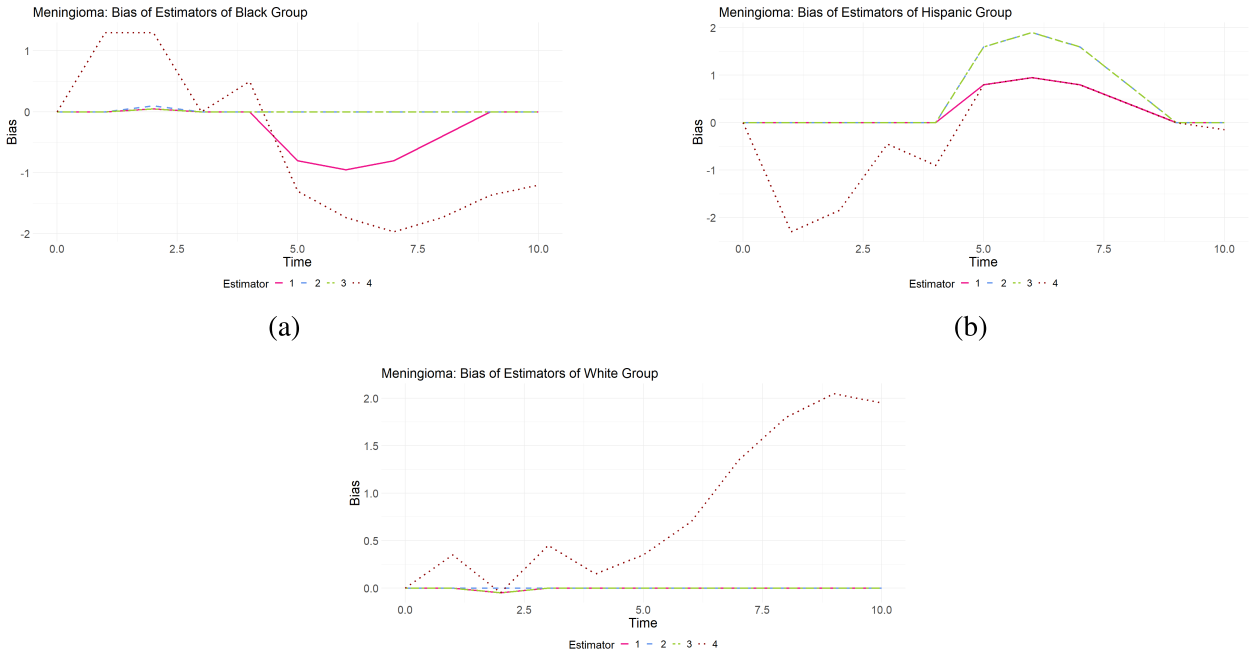
## 6.4 Meningioma of the Brain

| | Hispanic | Non-Hispanic Black | Non-Hispanic White |
|---|---|---|---|
| Estimator 1 | 0.21295 | 0.21318 | 0.00023 |
| Estimator 2 | 0.85182 | 0.00091 | 0.00000 |
| Estimator 3 | 0.85182 | 0.00023 | 0.00023 |
| Estimator 4 | 1.09909 | 1.68222 | 1.27545 |

Table 4: Table representing Mean Square Error (MSE) values for all estimators within the Meningioma Cancer category across the three populations: Non-Hispanic Black, Hispanic, and Non-Hispanic White.

For Meningioma estimator-based survival functions, see Figure 10 in the Appendix. The examination of Meningioma of the Brain data aligns with prior analyses of various cancer types, highlighting Estimator 4 as the most biased among the four estimators, evident in Figure 4. Notably, there is a pronounced increase in bias from Estimator 1 for the Black and Hispanic Groups, and from Estimator 3 for the Hispanic Group. Regarding MSE values, depicted in Figure 4, Estimator 1 exhibits the lowest MSE for the Hispanic Group, Estimator 3 for the Black Group, and Estimator 2 for the White Group. Despite these insights, determining the superior-performing model remains difficult in this instance.

The analysis of Meningioma of the Brain further underscores the variability in estimator performance. Estimator 4 stands out as the most biased, yet discerning the overall best-performing model is difficult.



(a)

(b)

(c)

Figure 4: Visual representation of bias trends for all estimators within the Meningioma category across distinct populations, including (a) Non-Hispanic Black, (b) Hispanic, and (c) Non-Hispanic White groups.

## 6.5 Mesothelioma

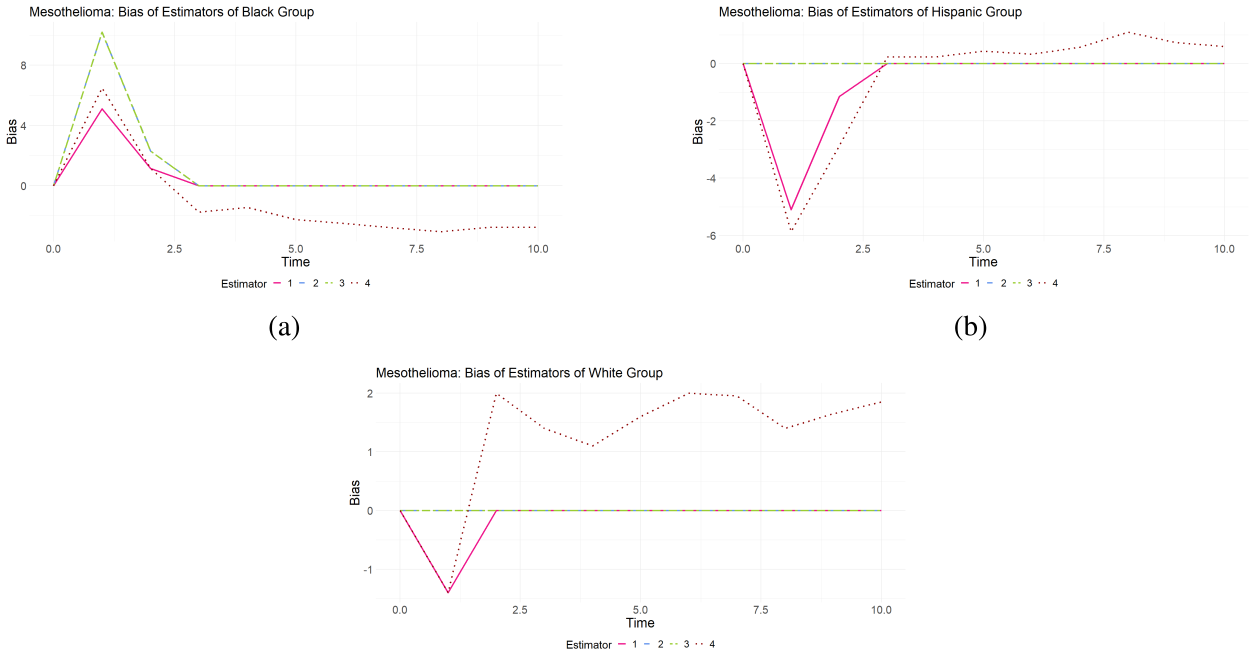
| | Hispanic | Non-Hispanic Black | Non-Hispanic White |
|---|---|---|---|
| Estimator 1 | 2.48477 | 2.48477 | 0.17818 |
| Estimator 2 | 0.00000 | 9.93909 | 0.00000 |
| Estimator 3 | 0.00000 | 9.93909 | 0.00000 |
| Estimator 4 | 4.13384 | 8.39250 | 2.50886 |

Table 5: Table representing Mean Square Error (MSE) values for all estimators within the Mesothelioma category across the three populations: Non-Hispanic Black, Hispanic, and Non-Hispanic White.

The examination of Mesothelioma case results, as depicted in Figure 5, introduces intriguing dynamics. For Mesothelioma estimator-based survival functions, see Figure 11 in the Appendix. There is a notable surge in bias from Estimator 1 for all three racial groups, accompanied by a similar trend from Estimators 2 and 3 for the Black Group. In contrast, Estimators 2 and 3 maintain a consistently bias-free profile for the White and Hispanic groups.

Regarding Mean Squared Error (MSE) values highlighted in Table 5, both Estimator 2 and Estimator 3 exhibit an MSE of 0 for the Hispanic and White Groups, while Estimator 1 claims the lowest MSE for the Non-Hispanic Black Group. Meanwhile, Estimator 4 consistently registers the highest MSE values across all racial categories.

Mesothelioma case results introduce intriguing complexities, with Estimator 1 exhibiting a notable bias surge. However, Estimators 2 and 3 maintain consistent bias-free profiles for certain demographic subsets.



(a)

(b)

(c)

Figure 5: Visual representation of bias trends for all estimators within the Mesothelioma category across distinct populations, including (a) Non-Hispanic Black, (b) Hispanic, and (c) Non-Hispanic White groups.

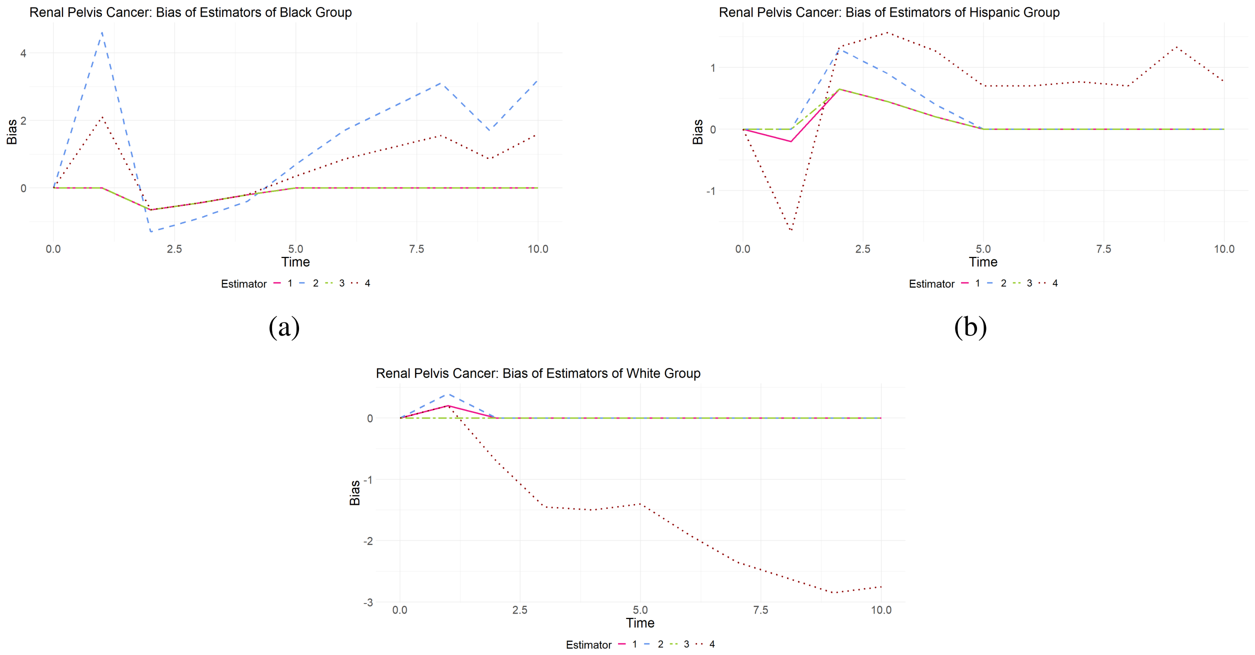
## 6.6 Renal Pelvis Cancer

| | Hispanic | Non-Hispanic Black | Non-Hispanic White |
|---|---|---|---|
| Estimator 1 | 0.06409 | 0.06045 | 0.00364 |
| Estimator 2 | 0.24182 | 5.06364 | 0.01455 |
| Estimator 3 | 0.06045 | 0.06045 | 0.00000 |
| Estimator 4 | 1.18525 | 1.18591 | 3.49273 |

Table 6: Table representing Mean Square Error (MSE) values for all estimators within the Renal Pelvis Cancer category across the three populations: Non-Hispanic Black, Hispanic, and Non-Hispanic White.

For Renal Pelvis Cancer estimator-based survival functions, see Figure 12 in the Appendix. In Table 6, we observe that Estimator 3 exhibits the lowest Mean Squared Error (MSE) for the Hispanic Group, while Estimators 1 and 3 share the lowest MSE for the Black Group. Additionally, Estimator 3 attains the lowest MSE value for the White Group. An intriguing result emerges in this scenario - Estimator 4, while not securing the highest MSE values across all categories, still falls short of being considered a top-performing model.

Regarding bias, as illustrated in Figure 6, Estimator 4 still is remarkably biased. Conversely, Estimators 1, 2, and 3 exhibit biases that hover relatively close to zero throughout the duration of the White Group analysis. Notably, Estimator 2 displays pronounced bias for the Black Group and experiences a substantial spike in bias for the Hispanic Group, introducing further nuances to the interpretation of model performance.



(a)

(b)

(c)

Figure 6: Visual representation of bias trends for all estimators within the Renal Pelvis Cancer category across distinct populations, including (a) Non-Hispanic Black, (b) Hispanic, and (c) Non-Hispanic White groups.

# 7 Conclusion

In conclusion, the comparative analysis of the four estimators - Estimator 1, Estimator 2, Estimator 3, and Estimator 4 - has provided valuable insights into their performances across different cancer types. Notably, Estimator 2 and Estimator 3 consistently demonstrated good performance, showcasing the lowest Mean Squared Error (MSE) values and minimal bias across multiple cancer categories.

An intriguing observation arises from the performance of Estimator 2 and Estimator 3, which, despite being the simpler estimators in the set, outperformed their more complex counterparts. This result highlights the importance of considering the trade-off between complexity and performance in selecting estimators for survival analysis in cancer research.

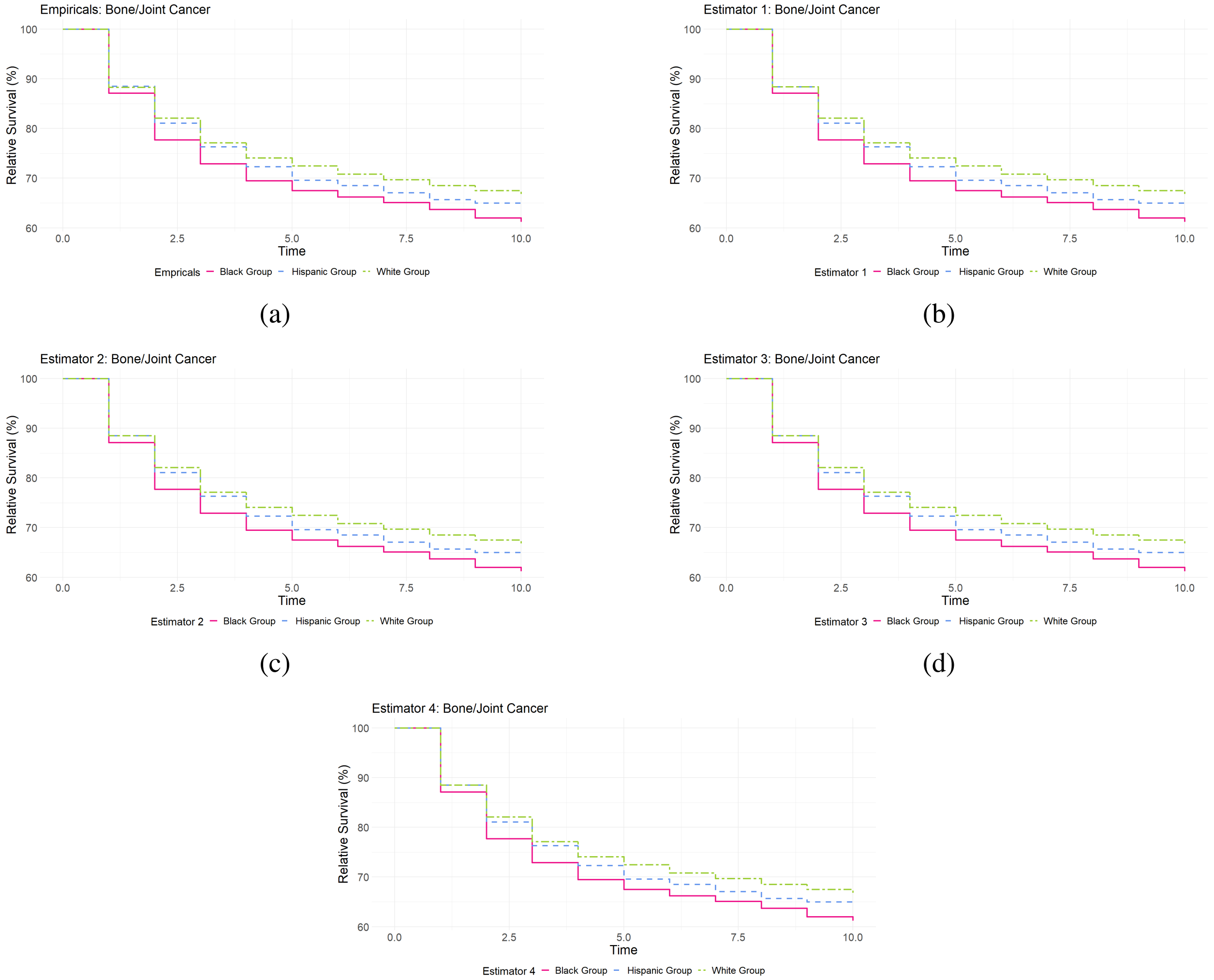
Furthermore, the performance of Estimator 2 in this study aligns with its consistent performance in preliminary research involving simulations at RUSIS@IU, underscoring its reliability and robustness across real-world data and simulated scenarios.

However, the comparative analysis also reveals the challenges in definitively declaring a single superior estimator across diverse cancer types. The nuanced performances and intricate dynamics observed underscore the need for further exploration, accounting for various factors influencing estimator performance. This study contributes to understanding estimator behaviors in cancer research, and could hold a role in refining future methodologies in the field.

While these results provide valuable insights, it is crucial to acknowledge the necessity for additional research in this area. The complexities of cancer data necessitate a comprehensive exploration of various factors influencing estimator performance, including dataset characteristics, sample size, and specific cancer types. Additionally, an in-depth investigation into the underlying mechanisms contributing to the superior performance of Estimator 2 could unveil novel avenues for refining survival analysis methodologies.

Ultimately, this study adds to the growing body of knowledge in survival analysis for cancer research and hopefully serves as a catalyst for future investigations. The findings of this study set the stage for more nuanced and sophisticated approaches to survival estimation, with the overarching goal of improving prognostic accuracy and advancing our understanding of cancer outcomes.

# Appendix



(a)

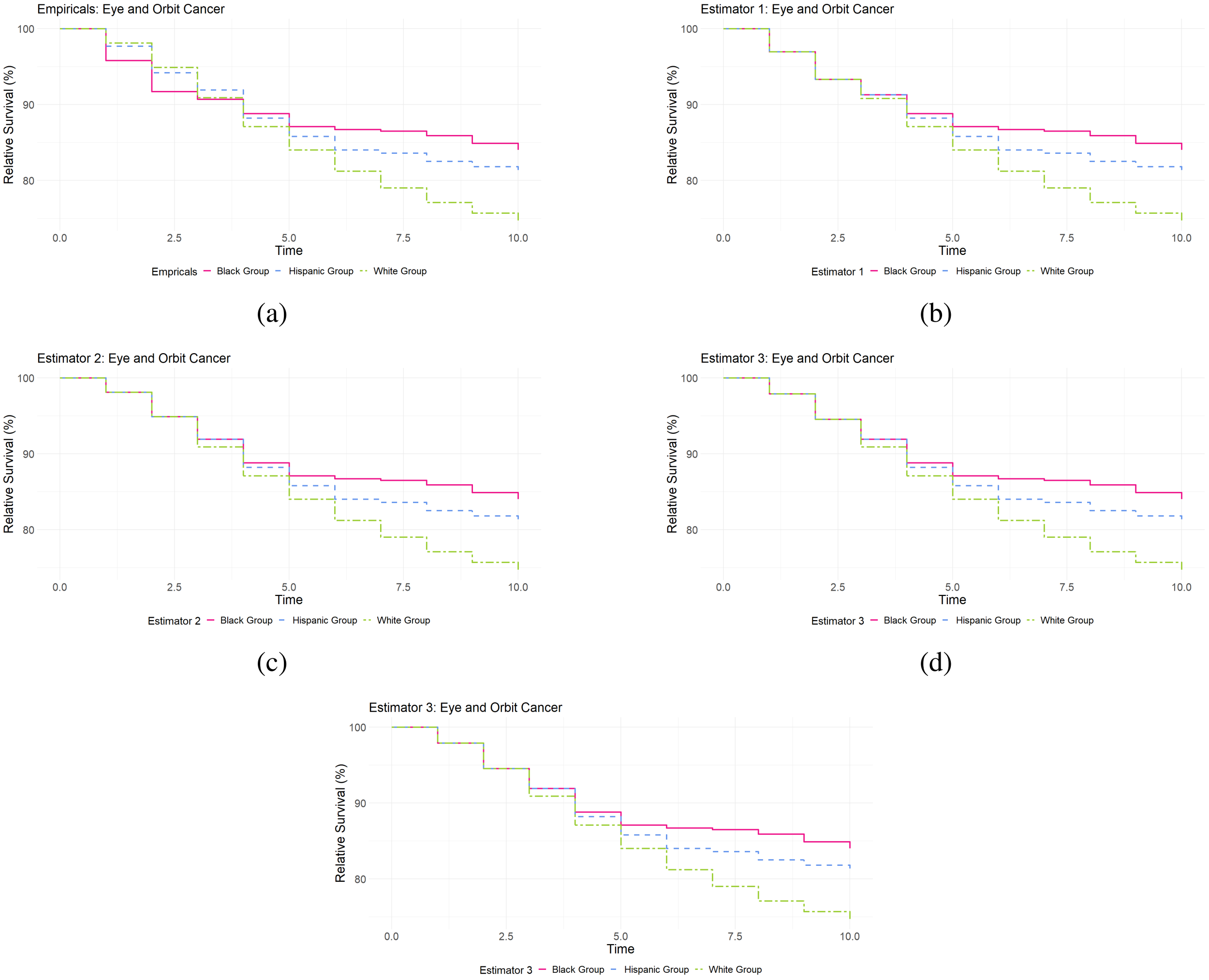
(b)

(c)

(d)

(e)

Figure 7: Visual representation of (a) the empirical survival functions from the raw data for the Bone/Joint Cancer case, and the new survival functions created after (b) Estimators 1, (c) 2, (d) 3, and (e) 4 were applied.



(a)

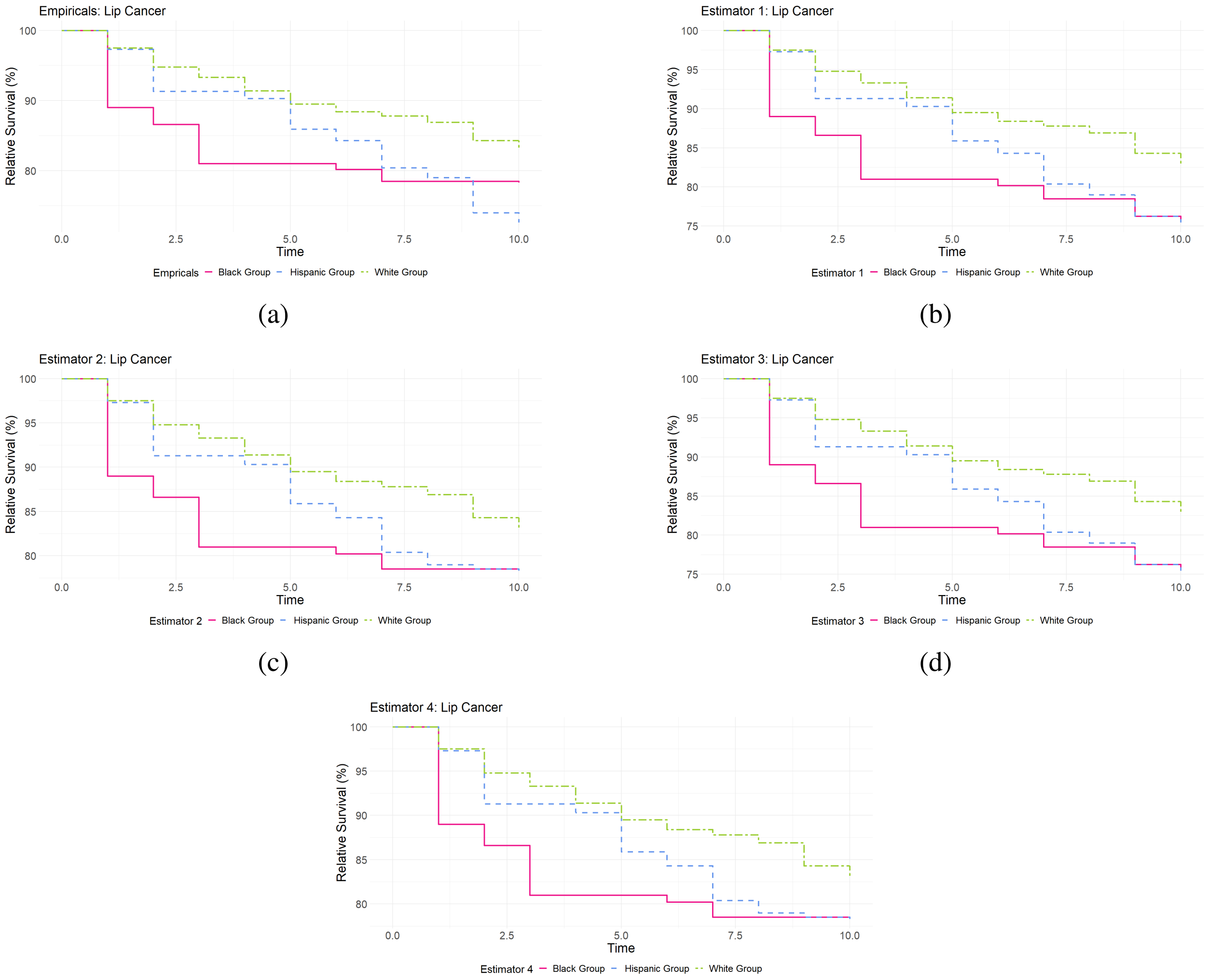
(b)

(c)

(d)

(e)

Figure 8: Visual representation of (a) the empirical survival functions from the raw data for the Eye/Orbit Cancer case, and the new survival functions created after (b) Estimators 1, (c) 2, (d) 3, and (e) 4 were applied.



(a)

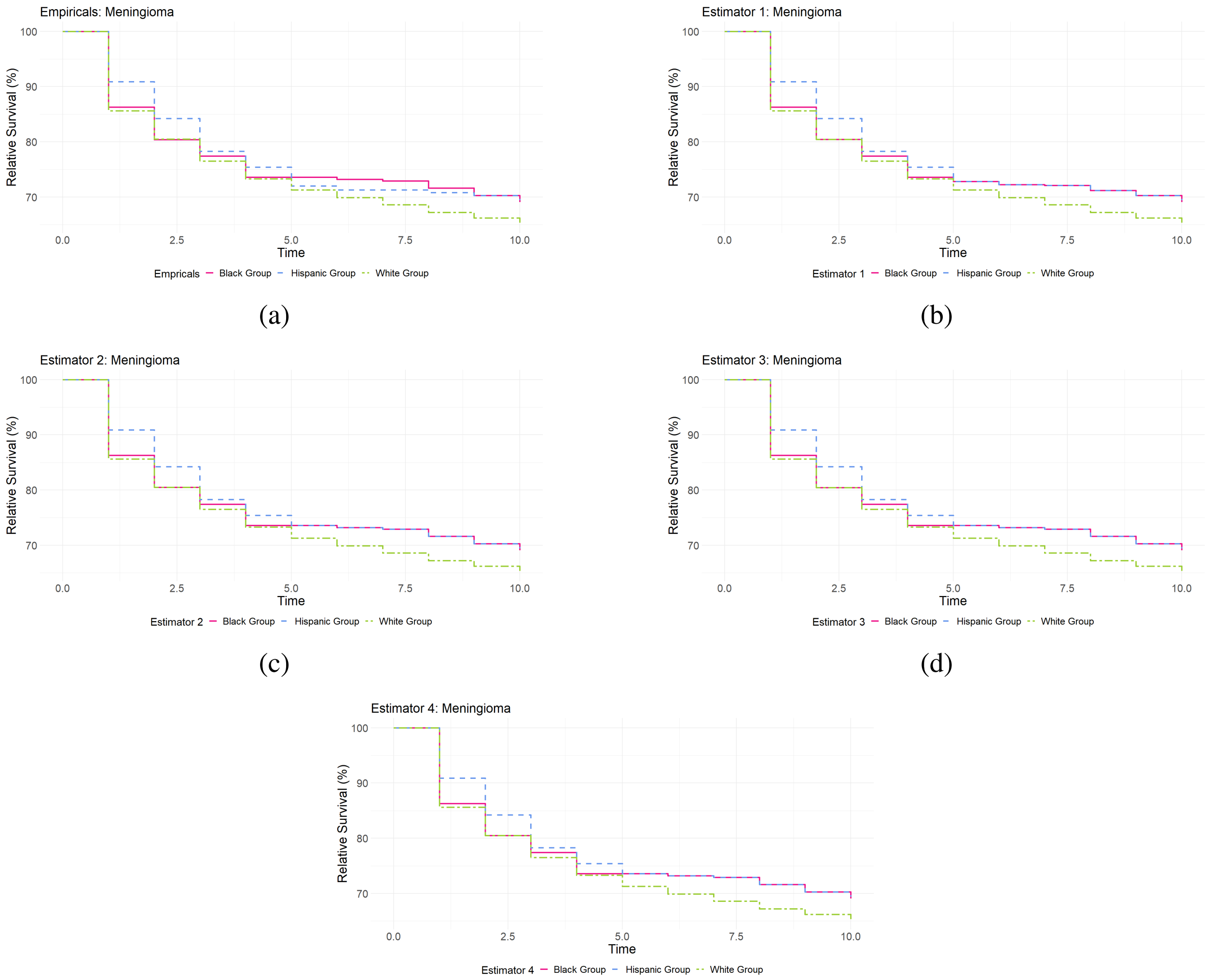
(b)

(c)

(d)

(e)

Figure 9: Visual representation of (a) the empirical survival functions from the raw data for the Lip Cancer case, and the new survival functions created after (b) Estimators 1, (c) 2, (d) 3, and (e) 4 were applied.



(a)

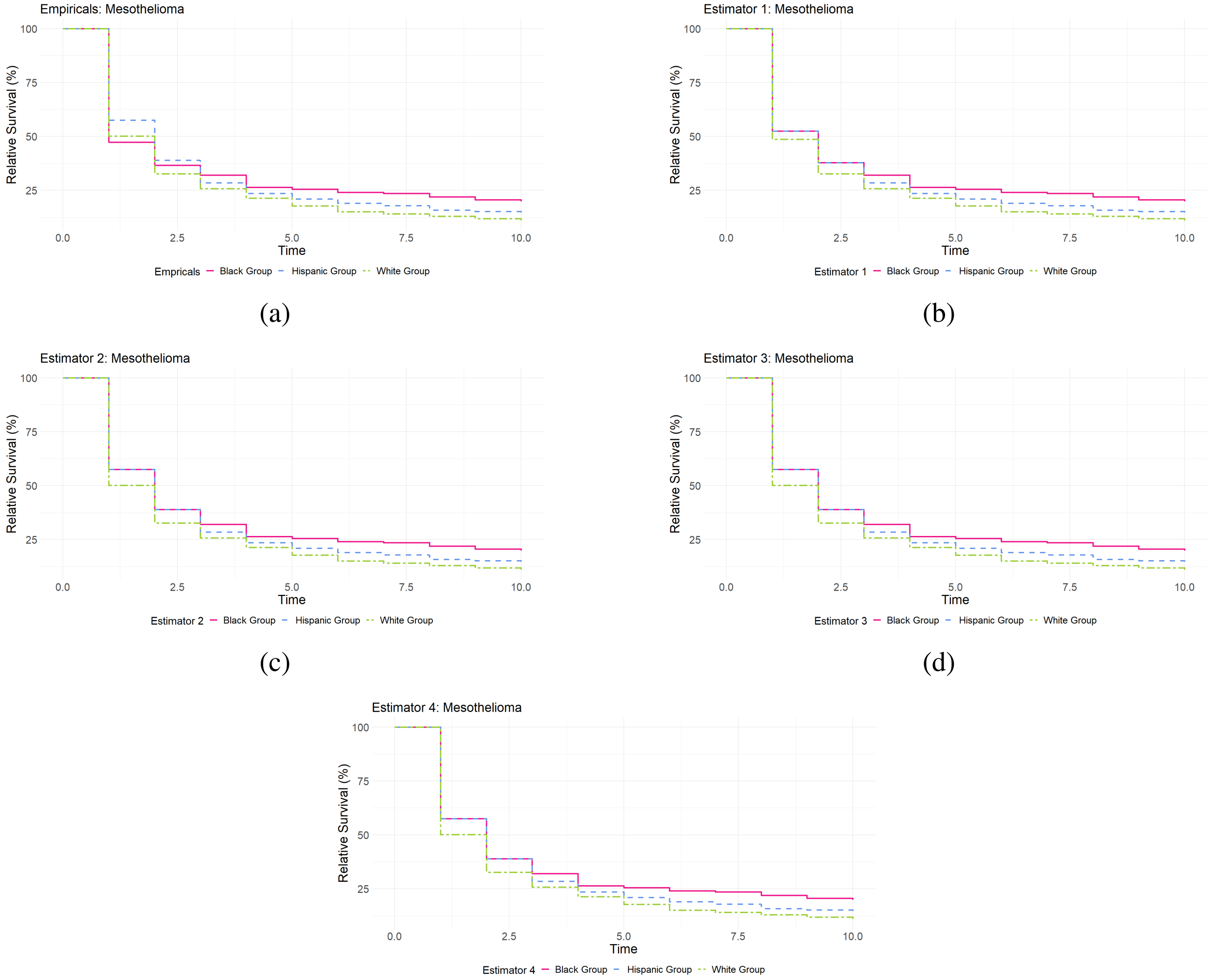
(b)

(c)

(d)

(e)

Figure 10: Visual representation of (a) the empirical survival functions from the raw data for the Meningioma case, and the new survival functions created after (b) Estimators 1, (c) 2, (d) 3, and (e) 4 were applied.



(a)

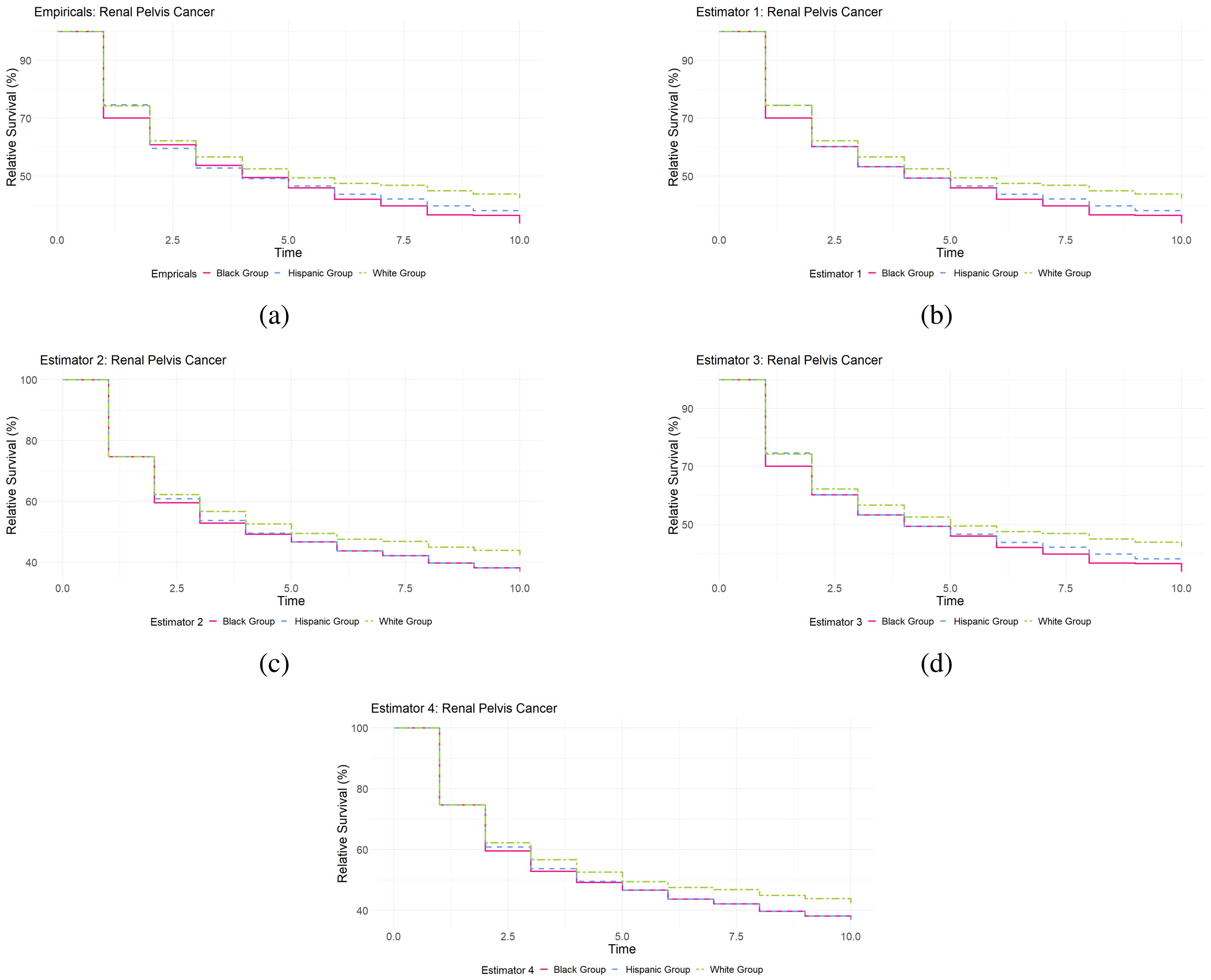
(b)

(c)

(d)

(e)

Figure 11: Visual representation of (a) the empirical survival functions from the raw data for the Mesothelioma case, and the new survival functions created after (b) Estimators 1, (c) 2, (d) 3, and (e) 4 were applied.



(a)

(b)

(c)

(d)

(e)

Figure 12: Visual representation of (a) the empirical survival functions from the raw data for the Renal Pelvis Cancer case, and the new survival functions created after (b) Estimators 1, (c) 2, (d) 3, and (e) 4 were applied.

# References

Arcones, M. A., Kvam, P. H., & Samaniego, F. J. (2002). Nonparametric estimation of a distribution subject to a stochastic precedence constraint. *Journal of the American Statistical Association*, *97*(457), 170–182. https://doi.org/10.1198/016214502753479310

Barmi, H. E., & Mukerjee, H. (2005). Inferences under a stochastic ordering constraint: The k-sample case. *Journal of the American Statistical Association*, *100*, 252–261. https://doi.org/10.1198/016214504000000764

El Barmi, H. (2017). Testing for uniform stochastic ordering via empirical likelihood under right censoring. *Statistica Sinica*, 645–664. https://www.jstor.org/stable/26383294

Jiménez, J. R., & Barmi, H. E. (2003). Estimation of distribution functions under second order stochastic dominance.

Lo, S.-H. (1987). Estimation of distribution functions under order restrictions. *Statistics & Risk Modeling*, *5*(3-4), 251–262. https://doi.org/10.1524/strm.1987.5.34.251

Oakes, D. (2000). Survival analysis. *Journal of the American Statistical Association*, *95*(449), 282–285. https://doi.org/10.2307/2669547

Park, Y., Kalbfleisch, J., & Taylor, J. (2012). Constrained nonparametric maximum likelihood estimation of stochastically ordered survivor functions. *Canadian Journal of Statistics*, *40*, 22–39. https://doi.org/10.1002/cjs

Park, Y., Taylor, J. M. G., & Kalbfleisch, J. D. (2012). Pointwise nonparametric maximum likelihood estimator of stochastically ordered survivor functions. *Biometrika*, *99*(2), 327–343. https://doi.org/10.1093/biomet/ass006

Puri, P. S., & Singh, H. (1992). Estimation of a distribution function dominating stochastically a known distribution function. *Australian Journal of Statistics*, *34*, 31–38. https://doi.org/10.1111/j.1467-842X.1992.tb01040.x

Rojo, J. (2004). On the estimation of survival functions under a stochastic order constraint. *The First Erich L. Lehmann Symposium-Optimality*, *44*, 37–61. https://doi.org/10.1214/lnms/1215006764

Rojo, J., & Ma, Z. (1996). On the estimation of stochastically ordered survival functions. *J. Statist. Comput. Simul*, *55*. https://doi.org/10.1080/00949659608811745

Surveillance, epidemiology, and end results (seer) program. (2023). www.seer.cancer.gov