Stockholms universitet

Simulation Study of Cumulative Hazard Estimators Under Nonparametric, Cox, and Gradient Boosting Frameworks

Yones Mestar

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2026:8
Matematisk statistik
Maj 2026

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2026:8**
http://www.math.su.se

Simulation Study of Cumulative Hazard Estimators Under Nonparametric, Cox, and Gradient Boosting Frameworks

Yones Mestar*

May 2026

Abstract

This thesis compares nonparametric and semi–parametric estimators of the cumulative hazard function through simulation studies, both with and without random right–censoring. Within the nonparametric framework, the Nelson–Aalen estimator is compared to a log transformation of the Kaplan–Meier estimator. In the Cox proportional hazards framework, the Breslow estimator with regression coefficients estimated by partial likelihood is compared to the Nelson–Aalen estimator with known coefficients, across unstratified and stratified models with varying covariate complexity. Finally, the Breslow estimator is evaluated with and without the assistance of a Gradient Boosting Machine estimating the log relative risk function, under both linear and nonlinear covariate structures. Performance is assessed using bias, integrated squared error and squared error at the last observed survival time, across a range of baseline hazard shapes. The results show that Nelson–Aalen and Kaplan–Meier perform comparably with Nelson–Aalen generally achieving lower integrated squared error, while the Breslow estimator dominates Nelson–Aalen as soon as multiple covariates are present. The Gradient Boosting Machine improves the Breslow estimator, especially under nonlinear covariate structures where the Cox model is unsuitable, while offering moderate improvement under linear structures.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: mestar.yones@gmail.com. Supervisor: Leo Levenius & Ola Hössjer.

Sammanfattning

Denna uppsats jämför icke-parametriska och semiparametriska skattare av den kumulativa hasardfunktionen genom simuleringsstudier, både med och utan slumpmässig högercensurering. Inom det icke-parametriska ramverket jämförs Nelson–Aalen-skattaren med en logaritmisk transformation av Kaplan–Meier-skattaren. Inom Cox proportionella hasardmodell jämförs Breslow-skattaren med regressionskoefficienter skattade via partiell sannolikhet mot Nelson–Aalen-skattaren med kända koefficienter, för ostratifierade och stratifierade modeller med varierande kovariatkomplexitet. Slutligen utvärderas Breslow-skattaren med och utan stöd av en Gradient Boosting Machine som skattar log–relativ–riskfunktionen, under både linjära och icke–linjära kovariatsstrukturer. Prestanda mäts med hjälp av bias, integrated squared error och squared error vid den sista observerade överlevnadstidpunkten, för ett antal olika baslinjehasardfunktioner. Resultaten visar att Nelson–Aalen och Kaplan–Meier presterar jämförbart där Nelson–Aalen generellt uppnår lägre integrated squared error, medan Breslow-skattaren dominerar Nelson–Aalen så snart flera kovariater förekommer. Gradient Boosting Machine förbättrar Breslow-skattaren, speciellt under icke-linjära kovariatsstrukturer där Cox-modellen är olämplig och ger måttliga förbättringar under linjära strukturer.

Acknowledgements

I would like to express my gratitude toward Leo Levenius for reaching out with the idea for this thesis and for his support throughout the process. I also thank Ola Hössjer for his valuable guidance and feedback. Perplexity was used for finding sources on subjects to research and ChatGPT assisted with text revision.

Contents

1 Introduction

Survival analysis described in Aalen et al. (2008, pp. 1–2), is a branch of statistics concerned with modeling the time until an event of interest occurs such as death, equipment failure, or disease relapse. A central quantity in this framework is the hazard rate, which describes the instantaneous risk of the event occurring at a given time, conditional on survival up to that point. Estimating the cumulative hazard function from observed survival data is a fundamental objective in this field as it summarizes the accumulated risk over time and forms the basis for inference about survival probabilities and covariate effects. An important question is how different estimators of the cumulative hazard perform relative to one another across various models and scenarios.

In this thesis, we conduct simulations with samples both with and without random right–censoring, where the individual observations are assumed to be independent. Within the framework of nonparametric intensity processes for counting processes, the Nelson–Aalen estimator is compared to a transformation of the Kaplan–Meier estimator. Additionally, the Breslow estimator used for Cox's proportional hazard regression model with covariate coefficients estimated through partial likelihood is compared to the Nelson–Aalen estimator in the same model, but with known coefficients. These comparisons are also extended to stratified models. Finally, the Breslow estimator is combined with a Gradient Boosting Machine to estimate the log relative risk function nonparametrically, and compared against the standard Breslow–Cox estimator under both linear and nonlinear covariate structures.

The hazard rates considered range from simple constant and linearly increasing shapes to exponentially increasing, stepwise and decreasing forms, providing a broad assessment of estimator performance across different data generating mechanisms. The covariate structures range from a single continuous covariate to combinations of continuous and discrete covariates with and without interaction terms which allows examination of the effect of increasing covariate complexity on estimation accuracy.

2 Survival Analysis

2.1 Survival Function and Hazard Rate

The following definitions and derivation of the relationship between the survival function and the cumulative hazard follow the presentation in Aalen et al. (2008, pp. 5–6).

Let us have an event for a population, such as death. The survival function, $S(t)$ is the probability that the event, T for a certain individual within the population occurs after the time point $t > 0$, i.e.

$$S(t) := \mathbb{P}(T > t). \tag{1}$$

Take note that $S(0) = 1$ due to the fact that it is guaranteed for the event to occur after the start observation point $t = 0$.

Assume now a stochastic variable T, then the hazard rate is defined as

$$\alpha(t) := \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t \leq T < t + \Delta t \mid T \geq t). \tag{2}$$

Equation (2) also indicates that $\alpha(t)dt$ is the probability that an individual of the population experience the event of interest in the time interval $[t, t + dt)$ for small dt, given that the event hasn't occurred for the person before the time t.

Lastly the cumulative hazard rate is defined as

$$A(t) = \int_0^t \alpha(s)ds, \tag{3}$$

which represents the accumulated "amount of risk" of the event to occur at time t. Utilizing (3) together with (1) and (2) yields

$$A'(t) = \alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{S(t) - S(t + \Delta t)}{S(t)} = -\frac{S'(t)}{S(t)}. \tag{4}$$

Integrating (4) with $S(0) = 1$ in mind, we get

$$\int_0^t -\frac{S'(s)}{S(s)}ds = \int_0^t \alpha(s)ds \iff$$

$$-\ln(S(t)) = A(t), \tag{5}$$

which is equivalent to

$$S(t) = e^{-A(t)}. \tag{6}$$

2.2 Censoring

Censored survival times are incomplete observations. Incomplete observations on the right–hand side of the time scale are referred to right–censored. In such cases, we know that the individual has survived up to the time of censoring, but we have no information beyond that point. For instance, medical studies usually have determined time limits so follow–up is often limited to a fixed study period and individuals who do not experience the event of interest before the end of the study are therefore right–censored. These concepts are elaborated upon in Aalen et al. (2008, pp. 3–5).

Random censoring is a type of censoring described in Aalen et al. (2008, pp. 29–30), in which a stochastic variable C_i represents the censoring time for individual i. Their observed survival time is given by $\tilde{T}_i = \min(T_i, C_i)$ and if $C_i < T_i$, we flag that the individual has been right–censored.

Random censoring arises in several settings. For example, in clinical studies with staggered enrollment and a fixed end date, individuals who enter later may be censored at the study's end, effectively making the censoring random due to differing enrollment times. Another example is loss to follow-up, where participants leave the study at unpredictable times for reasons unrelated to the event of interest. Similarly, in reliability studies, units may be withdrawn from testing at random times due to external constraints, resulting in right–censored observations.

2.3 Nonparametric Intensity Process

The nonparametric intensity process that we focus on are discussed in detail in Aalen et al. (2008, p. 34).

In this thesis, we apply nonparametric intensity processes for counting processes in order to compare nonparametric estimators such as the Nelson–Aalen estimator. The general counting process $N(t)$ considered for comparisons has an intensity process in the form of

$$\lambda(t) = \alpha(t)Y(t), \tag{7}$$

where $\alpha(t)$ is the hazard rate and $Y(t)$ is the number of individuals in the population at risk of the event when the time approaches the point t or in other words "just before t".

2.4 The Nelson–Aalen Estimator

Assume that we have survival data of n individuals with or without right–censoring and that no observation is tied to the same time point such that the observed survival times are $0 < T_1 < T_2 < \ ... < T_k$, where $k \leq n$. Let now $N(t)$ be a counting process that tracks the number of individuals experiencing

the event in question and its intensity process is nonparametric with the form (7). Following Aalen et al. (2008 pp. 71–72), the Nelson–Aalen estimator of the cumulative hazard rate (3) is

$$\hat{A}(t) = \sum_{T_j \leq t} \frac{1}{Y(T_j)}, \tag{8}$$

Note that (8) only sums over observed survival events.

With the estimator

$$\hat{\sigma}^2 = \sum_{T_j \leq t} \frac{1}{Y(T_j)^2} \tag{9}$$

of the variance of the Nelson–Aalen estimator we obtain the standard $100(1-\alpha)\%$ confidence interval for $A(t)$,

$$\hat{A}(t) \ \pm \ z_{1-\alpha/2}\hat{\sigma} \ , \tag{10}$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution.

2.5 The Kaplan–Meier Estimator

Let us have the same assumptions recently described in section 2.4. The Kaplan–Meier estimator of the survival function (1) is

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{1}{Y(T_j)}\right). \tag{11}$$

This construction and its properties are discussed in detail in Aalen et al. (2008 pp. 90–91).

Recall the connection between the survival function and cumulative hazard rate (6). Using this transformation we obtain the following estimation of the cumulative hazard rate

$$\hat{A}(t) = -\ \ln\left(\hat{S}(t)\right). \tag{12}$$

The confidence interval for $A(t)$ is

$$\left[-\ \ln\left(\hat{S}(t) + z_{1-\alpha/2}\hat{\sigma}\right), -\ \ln\left(\hat{S}(t) - z_{1-\alpha/2}\hat{\sigma}\right)\right], \tag{13}$$

which is the result of applying the strictly monotone transformation (12) on the confidence interval on Nelson–Aalen (10). Note that the boundary points are reversed, since the transformation is strictly decreasing. The use of strictly monotone functions to obtain confidence intervals is discussed in Held & Sabanés Bové (2020 p. 63).

2.6 Cox Regression Model

Following the counting process framework for survival analysis presented in Aalen et al. (2008, pp. 131–134), let us consider a counting process $N_i(t)$ that tracks whether the event occurs for a specific individual i with the semi–parametric intensity process,

$$\lambda_i(t) = Y_i(t)\alpha_0(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t)), \tag{14}$$

where $Y_i(t)$ equals 1 if the event has not yet occured for individual i, just before time t, and 0 otherwise. The function $\alpha_0(t)$ denotes the baseline hazard rate for the population. The factor $r(\boldsymbol{\beta}, \mathbf{x}_i(t))$ is called the relative risk function, consisting of the individual's covariates as the vector $\mathbf{x}_i(t)$, which may be time–dependent (e.g. age) or –constant (e.g. gender) and their associated regression coefficients stored in the vector $\boldsymbol{\beta}$. The intensity process (14) is called proportional hazard.

The Cox model has an exponential relative risk function, $r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = e^{\boldsymbol{\beta}^T \mathbf{x}_i(t)}$ which is a linear combination of all covariates. Aggregating the counting processes of n individuals yields a population–level counting process

$$N(t) = \sum_{i=1}^{n} N_i(t), \tag{15}$$

whose intensity process is given by

$$\lambda(t) = \sum_{i=1}^{n} Y_i(t)\alpha_0(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = \alpha_0(t)\sum_{i=1}^{n} Y_i(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t)). \tag{16}$$

The Cox model is semi–parametric so in order to estimate $\boldsymbol{\beta}$, we apply the partial likelihood method as presented in Aalen et al. (2008, pp. 134–135). The intensity process (14) of individual i's counting process can be expressed as

$$\lambda_i(t) = \lambda(t)\pi(i \,|\, t), \tag{17}$$

where

$$\pi(i \,|\, t) = \frac{\lambda_i(t)}{\lambda(t)} = \frac{Y_i(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t))}{\sum_{l=1}^{n} Y_l(t)r(\boldsymbol{\beta}, \mathbf{x}_l(t))} \tag{18}$$

is the conditional probability of observing an event for individual i at time t given the history up to time t and that an event occurs at that time. This conditional probability does not depend on the baseline hazard rate $\alpha_0(t)$. Additionally, this method focuses on relative risks at event times, thus disregarding the information of regression coefficients $\boldsymbol{\beta}$ contained in the aggregated counting process $N(t)$.

Assume that we have a sample of k observed survival times with no ties in a sample of n survival times, e.g. $0 < T_1 < T_2 < \ ... < T_k$ and $k \leq n$. Then the

partial likelihood is

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \pi(i_j \,|\, T_j) = \prod_{j=1}^{k} \frac{Y_{i_j}(T_j) r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{l=1}^{n} Y_l(T_j) r(\boldsymbol{\beta}, \mathbf{x}_l(T_j))}, \tag{19}$$

where the index i_j corresponds to the individual that experienced the event at time T_j.

With (19), the maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}$ is obtained by maximizing $L(\boldsymbol{\beta})$ with respect of $\boldsymbol{\beta}$. This may also be done with the logarithmic partial likelihood

$$\begin{aligned} l(\boldsymbol{\beta}) &= \ln\left(\prod_{j=1}^{k} \pi(i_j \,|\, T_j)\right) = \sum_{j=1}^{k} \ln(\pi(i_j \,|\, T_j)) = \\ &= \sum_{j=1}^{k} \ln\left(\frac{Y_{i_j}(T_j) r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{l=1}^{n} Y_l(T_j) r(\boldsymbol{\beta}, \mathbf{x}_l(T_j))}\right). \end{aligned} \tag{20}$$

2.7 The Breslow Estimator

Let us assume a Cox model with intensity function

$$\lambda(t) = \alpha_0(t) \sum_{i=1}^{n} Y_i(t) r(\boldsymbol{\beta}, \mathbf{x}_i(t)),$$

and suppose that we want to estimate the cumulative baseline hazard rate $A_0(t)$. If the regression coefficients $\boldsymbol{\beta}$ were known, then $A_0(t)$ could be estimated with the Nelson–Aalen estimator

$$\hat{A}_0(t\,;\,\boldsymbol{\beta}) = \int_0^t \frac{dN(u)}{\sum_{l=1}^{n} Y_l(u) r(\boldsymbol{\beta}, \mathbf{x}_l(u))}, \tag{21}$$

where $dN(u)$ denotes the number of individuals experiencing the event at time u, which in our case always equals 0 or 1 due to the assumption of no tied events.

Since $\boldsymbol{\beta}$ is unknown, we replace it with by its maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ to obtain the Breslow estimator

$$\hat{A}_0(t\,;\,\hat{\boldsymbol{\beta}}) = \int_0^t \frac{dN(u)}{\sum_{l=1}^{n} Y_l(u) r(\hat{\boldsymbol{\beta}}, \mathbf{x}_l(u))} = \sum_{T_j < t} \frac{1}{\sum_{l=1}^{n} Y_l(T_j) r(\hat{\boldsymbol{\beta}}, \mathbf{x}_l(T_j))}, \tag{22}$$

as presented in Aalen et al. (2008, p. 141).

2.8 Stratified Models

In stratified Cox models which are presented and closely followed in Aalen et al. (2008 p. 148), the population is grouped into k strata such that the hazard rate for individual i in stratum s is

$$\alpha(t\,;\,\mathbf{x}_i) = \alpha_{s0}(t) r(\boldsymbol{\beta}, \mathbf{x}_i(t)). \tag{23}$$

This approach is useful in situations where the assumption of a common baseline hazard rate for the entire population is inappropriate. Further, the stratum membership may also be time dependent.

With the assumption of no ties for survival events, the regression coefficient $\boldsymbol{\beta}$ is estimated by maximizing the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{\mathcal{S}} \prod_{j=1}^{k_s} \frac{Y_{i_{sj}}(T_{sj}) r(\boldsymbol{\beta}, \mathbf{x}_{i_{sj}}(T_{sj}))}{\sum_{l=1}^{n_s} Y_l(T_{sj}) r(\boldsymbol{\beta}, \mathbf{x}_l(T_{sj}))} \tag{24}$$

where the observed survival times in stratum s are $0 < T_{s1} < T_{s2} < \ldots < T_{sk_s}$, k_s is the number of observed events, n_s is the number of individuals with $k_s \leq n_s$, index sl and i_{sj} denote, respectively, individual l and the individual experiencing the j–th event in stratum s. Finally, the Breslow estimator of the stratum–specific cumulative baseline hazard rate is

$$\hat{A}_{s0} = \sum_{T_{sj} < t} \frac{1}{\sum_{l=1}^{n} Y_{sl}(T_{sj}) r(\hat{\boldsymbol{\beta}}, \mathbf{x}_{sl}(T_{sj}))}. \tag{25}$$

If we had the true value of $\boldsymbol{\beta}$ we could apply the Nelson–Aalen estimator instead by replacing $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}$ in (25).

3 Decision Trees

We briefly describe decision trees following the overview given by GeeksforGeeks (2025, Decision Tree). A Decision Tree is a machine learning model with a hierarchical, tree–like structure. It predicts a target variable of a sample by starting at the root node, where a binary question is applied to the sample's variables to partition it into subsets. An example of a binary question could be whether the color is red or not. The tree then branches into child nodes depending on the outcome of the question. This process repeats at each child node and their observations, where a new binary question is evaluated on a selected variable to further split the data. This continues until the last node (also known as a leaf node) is reached, where the model finally assigns a predicted value to the target variable.

During a decision tree's training, an impurity measure is applied at each node to determine the optimal binary question (also referred to as split or threshold) to split the data, as in Yuksel & Aydede (2025, Ch. 13.1 & 13.4). A node with high impurity contains observations with highly varied values of the target variable, meaning the data is not separated well and a node with low impurity contains observations with similar values of the target variable, meaning the data is separated well. Common impurity measures include Gini impurity and entropy.

A concrete summary of pruning and overfitting is avaiable in GeeksforGeeks (2025, Decision Tree). A decision tree overfits when it grows too deep and starts memorizing the noise in the training data instead of learning general patterns that hold for new data. To prevent this, pruning is applied where we remove branches that contribute little to predictive performance, which reduces complexity and helps the tree generalize better to new data.

3.1 Regression Trees

We follow the presentation of regression trees in Yuksel & Aydede (2025, Ch. 13.4). A regression tree is a type of decision tree whose target variable is continuous. The impurity measure for a node v is typically the mean squared error (MSE) which is defined as,

$$\text{MSE}(v) = \frac{1}{M_v} \sum_{i \in v} (Z_i - \bar{Z}_v)^2, \tag{26}$$

where M_v denotes the number of observations in the node, Z_i is the value of the target variable for observation i and $\bar{Z}_v$ denotes the mean of target variable within node v. That way, low MSE indicates low impurity, making it an appropriate impurity measure for regression trees.

The tree evaluates the reduction in MSE that would result from potential splits in order to choose the best split for a given node. Assume we have a parent

node v that gets split into left and right child nodes v_L and v_R. Then the MSE reduction for the parent node v is

$$\Delta\text{MSE}(v) = \text{MSE}(v) - \frac{M_L}{M_v}\text{MSE}(v_L) - \frac{M_R}{M_v}\text{MSE}(v_R), \tag{27}$$

where M_L and M_R are the numbers of observations in the left and right child nodes, respectively.

For each node during training, the algorithm seeks the optimal split on the optimal variable that yields the greatest MSE reduction. Once the optimal split is determined, the node branches into its left and right children, and the process repeats recursively. A branch terminates when at least one stopping criterion is met, namely a minimum number of observations per leaf node, a maximum tree depth, or insufficient MSE reduction. These stopping criteria can be adjusted according to one's preferences.

For continuous and discrete predictor variables, a split is defined by a single threshold value. For instance, consider a variable that takes natural numbers or continuous values in the interval $[0, 1]$. In the former case, a split on a node could be defined such that an observation is sent to the left child if its variable value is less than 3, and to the right child otherwise. Binary or dummy variables are simply split according to their two unique values.

The predicted value of the target variable for a leaf node v is the mean of the target variable over all observations in that leaf,

$$\hat{Z}_v = \bar{Z}_v = \frac{1}{M_v}\sum_{i \in v} Z_i. \tag{28}$$

Pruning a tree's leaf nodes and branches is an important concept explained in Yuksel & Aydede (2025, Ch. 13.2). The fully grown tree is pruned using the cost–complexity function

$$\kappa_\alpha(\mathcal{L}) = \sum_{v \in \mathcal{L}} M_v \cdot \text{MSE}(v) + \alpha \cdot |\mathcal{L}|, \tag{29}$$

where $\mathcal{L}$ is the set of all leaves and M_v denotes the number of observations in leaf node v. Larger values of α impose a greater penalty on tree complexity, thereby incentivizing trees with fewer leaves. For a given α, the optimal subtree minimizes $\kappa_\alpha(\mathcal{L})$. As α increases, the penalty on tree size grows and successively simpler subtrees become optimal, generating a nested sequence of subtrees $\mathcal{T}_1 \supset \mathcal{T}_2 \supset \cdots \supset \mathcal{T}_m$ with corresponding penalty values $\alpha_1 < \alpha_2 < \cdots < \alpha_m$, where $\mathcal{T}_1$ is the fully grown tree and $\mathcal{T}_m$ is the trivial tree consisting of only the root node.

The optimal penalty parameter α^* is selected through K-fold cross–validation, described in Yuksel & Aydede (2025, Ch. 13.4). The full sample is partitioned

into K subsets and the tree is trained on $K-1$ subsets and and tested on the remaining subset. This process is repeated K times, each time using a different subset for validation. For each candidate value of α, the cross–validation error is computed as the average MSE across all K folds, and the optimal value is then chosen as

$$\alpha^* = \arg \min_{\alpha} \text{CV}_{\text{ERROR}}(\alpha). \tag{30}$$

3.2 Ensemble Learning and Boosting

Ensemble Learning, elaborated in detail in Yuksel & Aydede (2025, Ch. 14), combines multiple models to improve prediction accuracy and reduce overfitting. The intuition is analogous to aggregating opinions from a crowd rather than relying on a single individual. A single model may discard weak predictors entirely, whereas ensemble methods retain them, allowing their collective contribution to inform the final prediction. The most common ensemble methods are bagging, random forests, and boosting.
Boosting is an ensemble method that trains models (or regression trees in our case) sequentially, where each model focuses on correcting the errors of its predecessor, as presented in Natekin & Knoll (2013). These models, commonly referred to as weak learners are relatively simple. By combining many such learners, boosting yields a strong learner whose collective strengths outperform any individual model.

The algorithm begins by training an initial weak learner on the full sample. Subsequent weak learners are then fitted to the residuals of the current ensemble, progressively reducing the errors that remain. This sequential correction primarily reduces bias, though it may also influence variance.

3.3 Gradient Boosting Machines

Gradient Boosting Machine (GBM) is a boosting algorithm presented in Natekin & Knoll (2013), where each new weak learner is trained to minimize a differentiable loss function $L(z_i, \hat{z}_i)$ with respect to predictions using gradient descent. Here z_i is the value of the target variable for observation i, whereas $\hat{z}_i$ is the current ensemble learner's prediction of this target variable. The square error loss function $L(z_i, \hat{z}_i) = \frac{1}{2}(z_i - \hat{z}_i)^2$ is frequently used for GBMs.

The algorithm begins by training an initial model on a sample of n observations and computing the gradient with respect to current predictions,

$$\nabla L = \left(\frac{\delta L(z_1, \hat{z}_1)}{\delta \hat{z}_1}, \frac{\delta L(z_2, \hat{z}_2)}{\delta \hat{z}_2}, \cdots, \frac{\delta L(z_n, \hat{z}_n)}{\delta \hat{z}_n} \right). \tag{31}$$

The gradient points in the direction of the steepest increase in L, and therefore the negative gradient points in the direction of the steepest decrease. Under squared error loss, the negative gradient is simply a vector of the residuals

$z_i - \hat{z}_i$, meaning each new weak learner is fitted to the residuals of the current ensemble. The steps of the algorithm are summarized in Table 1.

Table 1: Steps of the GBM algorithm.

Step	Description
1	Initialize the ensemble with a constant prediction $\hat{Z}_{i0}$ that minimizes L.
2	For each iteration $a = 1, 2, \ldots, M$:
3	Compute the negative gradient for each observation $i = 1, \ldots, n$: $r_{ia} = -\frac{\delta L(z_i, \hat{z}_i)}{\delta \hat{z}_i}$
4	Fit a new weak learner h_j to the pseudo–residuals $\{r_{ia}\}_{i=1}^n$, yielding predicted values $\hat{r}_{ia} = h_a(\mathbf{x}_i)$.
5	Update the ensemble with learning rate η: $\hat{Z}_i \leftarrow \hat{Z}_i + \eta \cdot \hat{r}_{ia}$.
6	Repeat Steps 3–5 until $a = M$ or early stopping is triggered.

Following Table 1, a new weak learner is then trained on the negative gradient of the previous ensemble and its predictions are added to the ensemble proportionally, progressively minimizing L. This process is repeated until a pre–specified number of iterations $\mathcal{M}$ is reached or early stopping via cross–validation is triggered.

Suppose that we have trained $\mathcal{M}$ weak learners with this process, the ensemble's prediction of Z_i is

$$\hat{Z}_i = \hat{Z}_{i0} + \eta\hat{r}_{i1} + \eta\hat{r}_{i2} + \cdots + \eta\hat{r}_{i\mathcal{M}}, \tag{32}$$

where $\hat{Z}_{i0}$ denotes the initial model's prediction of Z_i, $\hat{r}_{ia}$ is the $j+1$–th weak learner's predicted negative gradient at iteration a for observation i, and η denotes the learning rate, which assumes a value in the interval $]0, 1]$. A small learning rate reduces the contribution of each new weak learner, lowering the risk of overfitting at the cost of requiring more weak learners to achieve a strong performance. Conversely, a large learning rate increases the impact of the new weak learner but raises the risk of overfitting.

3.4 GBM for Cox Proportional Hazards Models

To apply the GBM algorithm within the Cox proportional hazards framework, the squared error loss is not suitable. This is due to right–censoring, which prevents direct computation of residuals, and because the target quantity is the log relative risk $f(\mathbf{x}_i) = \ln r(\boldsymbol{\beta}, \mathbf{x}_i(t))$ rather than an observed continuous response. Instead, we use the negative Cox partial log–likelihood as the loss function, thus minimizing the negative partial log–likelihood is equivalent to maximizing the partial likelihood (19), which means the GBM is implicitly seeking the same objective as Cox's partial likelihood estimation, but without restricting $(\mathbf{x}_i)$ to be linear in the covariates.

Assume a sample of n individuals with $k \leq n$ observed events at ordered times $0 < T_1 < T_2 < \cdots < T_k$. The loss function is defined as

$$L = -\sum_{j=1}^{k} \left(\hat{f}(\mathbf{x}_{i_j}) - \log \sum_{l=1}^{n} Y_l(T_j) e^{\hat{f}(\mathbf{x}_l)} \right), \tag{33}$$

where i_j refers to the individual who experiences event j at time T_j.

To apply gradient boosting, we compute the negative gradient of L with respect to the predictions $\hat{f}(\mathbf{x}_i)$. Differentiating (33) gives

$$\frac{\partial L}{\partial \hat{f}(\mathbf{x}_i)} = -\sum_{j=1}^{k} \left[\frac{\partial}{\partial \hat{f}(\mathbf{x}_i)} \hat{f}(\mathbf{x}_{i_j}) - \frac{\partial}{\partial \hat{f}(\mathbf{x}_i)} \log \sum_{l=1}^{n} Y_l(T_j) e^{\hat{f}(\mathbf{x}_l)} \right]. \tag{34}$$

The first term evaluates to

$$\frac{\partial}{\partial \hat{f}(\mathbf{x}_i)} \hat{f}(\mathbf{x}_{i_j}) = \mathbf{1}(i = i_j), \tag{35}$$

where $\mathbf{1}(i = i_j)$ is an indicator function equal to 1 if individual i is the one experiencing the event at time T_j, and 0 otherwise.

The second term becomes

$$\frac{\partial}{\partial \hat{f}(\mathbf{x}_i)} \log \sum_{l=1}^{n} Y_l(T_j) e^{\hat{f}(\mathbf{x}_l)} = \frac{Y_i(T_j) e^{\hat{f}(\mathbf{x}_i)}}{\sum_{l=1}^{n} Y_l(T_j) e^{\hat{f}(\mathbf{x}_l)}}. \tag{36}$$

Combining these results, the negative gradient (pseudo-residual) for observation i at iteration a is

$$r_{ia} = -\frac{\partial L}{\partial \hat{f}(\mathbf{x}_i)} = \sum_{j=1}^{k} \left[\mathbf{1}(i = i_j) - \frac{Y_i(T_j) e^{\hat{f}(\mathbf{x}_i)}}{\sum_{l=1}^{n} Y_l(T_j) e^{\hat{f}(\mathbf{x}_l)}} \right]. \tag{37}$$

These pseudo–residuals represent the difference between the observed events and the model's expected contribution at each risk set.

Following the GBM procedure, at each iteration a, a regression tree h_a is fitted to the pseudo-residuals $\{r_{ia}\}_{i=1}^{n}$, producing predictions $\hat{r}_{ia} = h_a(\mathbf{x}_i)$. The model is then updated as

$$\hat{f}(\mathbf{x}_i) \leftarrow \hat{f}(\mathbf{x}_i) + \eta \hat{r}_{ia}. \tag{38}$$

After $\mathcal{M}$ iterations or when an early stop criterion is triggered, the ensemble approximates the log relative risk function $f(\mathbf{x}_i)$.

4 Simulation Study

We simulate samples of 1,000 survival times T_i, $1 \leq i \leq 1000 = n$, with all observations generated independently. For each (baseline) hazard rate, samples are generated without right–censoring and the estimators are compared. The same comparison is then repeated using the same samples but with random right–censoring, where the censoring times are also independent of the survival times. All simulations and analyses are performed in R.

The results of the simulation studies are presented using graphical summaries and performance metrics such as the mean bias,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{A}(t_i) - A(t_i)\right), \tag{39}$$

hereafter referred to as bias. The bias measures the average deviation of the estimated cumulative hazard from the true one, indicating whether the estimator systematically over- or underestimates the hazard.

In addition, the Integrates Squared Error (ISE) is used to assess overall estimation accuracy, as it penalizes large point–wise deviations more heavily than small ones. According to Emergent Mind (2025) the ISE with respect to the true cumulative hazard and its estimate is:

$$\mathrm{ISE}(\hat{f}) = \int_0^T \left(\hat{A}(t) - A(t)\right)^2 dt, \tag{40}$$

where T denotes the latest observed survival time in the sample. Since $\hat{A}(t)$ is discrete, we estimate ISE using Riemann sums:

$$\hat{\mathrm{ISE}}(\hat{f}) = \sum_{i=1}^{n}\left(\hat{A}(t_i) - A(t_i)\right)^2 \Delta t_i, \tag{41}$$

where Δt_i denotes $t_i - t_{i-1}$ and t_i is the survival time of observation i, assuming that the survival times have been ordered so that $0 < t_1 < \ldots < t_n = T$. Note that for $i = 0$, t_{i-1} is set to 0.

Since estimation uncertainty increases at later survival times due to fewer individuals remaining at risk, the Squared Error (SE) $(\hat{A}(t_n) - A(t_n))^2$ is evaluated for the last survival time t_n, since the squared error metric penalizes large deviations.

4.1 Nelson–Aalen vs. Kaplan–Meier

In the comparison of Nelson–Aalen against Kaplan–Meier, we are in the framework of nonparametric intensity processes and the cumulative hazard rates of the following hazard rates, defined for $t > 0$, will be estimated,

$$
\begin{aligned}
\alpha_1(t) &= 0.2, \\
\alpha_2(t) &= 0.2t, \\
\alpha_3(t) &= 0.2e^{0.1t}, \\
\alpha_4(t) &= \begin{cases} 0.1, & 0 < t \leq 10 \\ 0.3, & 10 < t \leq 20 \\ 0.15, & 20 < t \end{cases}, \\
\alpha_5(t) &= 0.2t^{-0.5}.
\end{aligned}
$$

Thus, we consider constant, linearly increasing, exponentially increasing, stepwise and decreasing hazard rates.

In order to simulate survival events from invertible cumulative hazard rates we extract samples from the uniform distribution, $U \sim \text{Uniform}(0, 1)$. This is justified by the probability integral transform presented in Casella & Berger (2002, p. 54). If X is a continuous random variable with CDF $F_X(x)$, then $U = F_X(x) \sim \text{Uniform}(0, 1)$, since for $u \in [0, 1]$,

$$
\mathbb{P}(F_X(X) \leq u) = \mathbb{P}(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u, \tag{42}
$$

which is exactly the CDF of the Uniform$(0, 1)$ distribution. Conversely, if $U \sim \text{Uniform}(0, 1)$, then $X = F_X^{-1}(U)$ has CDF F_X. Since the survival function satisfies $S(t) = \mathbb{P}(T > t) = 1 - F_T(t)$, and $1 - U \sim \text{Uniform}(0, 1)$ whenever $U \sim \text{Uniform}(0, 1)$, the transform applies equally to S. Evaluating the survival function at T yields

$$
\begin{aligned}
\mathbb{P}(S(T) \leq u) = \mathbb{P}(T \geq S^{-1}(u)) = 1 - F_T(S^{-1}(u)) = \\
= 1 - (1 - S(S^{-1}(u))) = u
\end{aligned} \tag{43}
$$

so that $U = S(T) \sim \text{Uniform}(0, 1)$. Using the connection between the survival function and the cumulative hazard rate (6) we set

$$
U = S(T) = e^{-A(T)} \iff -\ln(U) = A(T) \iff T = A^{-1}(-\ln(U)), \tag{44}
$$

where the invertibility of A guarantees a unique solution T for every draw of U.

For the stepwise hazard $\alpha_4(t)$, the cumulative hazard is

$$
A_4(t) = \begin{cases} 0.1t, & 0 < t \leq 10, \\ 1 + 0.3(t - 10), & 10 < t \leq 20, \\ 4 + 0.15(t - 20), & 20 < t, \end{cases}
$$

with breakpoints $A_4(10) = 1$ and $A_4(20) = 4$. Given a draw $E = -\ln(U)$, the inverse is obtained by locating which interval E falls in and inverting the

corresponding linear piece:

$$A_4^{-1}(e) = \begin{cases} e/0.1, & 0 < e \leq 1, \\ 10 + (e-1)/0.3, & 1 < e \leq 4, \\ 20 + (e-4)/0.15, & e > 4. \end{cases}$$

Since each piece of A_4 is strictly increasing and continuous, the inverse is well-defined and unique across all intervals.

The last observation of the uncensored sample is removed since the Kaplan–Meier survival estimate is 0 for that observation, meaning that the transformation (4) will become infinite. The censoring times are generated independently from an exponential distribution Exp(0.1) with intensity 0.1 and expected value 10.

4.2 Nelson–Aalen vs. Breslow–Cox

The Nelson–Aalen and Breslow estimators of the baseline cumulative hazard function, as defined in (21) and (22) respectively, are compared for Cox models with one continuous covariate that is standard normally distributed, one continuous and one discrete covariate where the discrete covariate is $\text{Bin}(1, 0.5)$ distributed, and lastly both covariates together with their interaction. The covariates are independent and their coefficients are set to $\beta_1 = 0.5$, $\beta_2 = -0.7$, and $\beta_3 = 0.2$, respectively. The same comparison is done on a stratified Cox model with the same continuous and discrete covariates but without their interaction, with coefficients $\beta_1 = 1.5$ and $\beta_2 = 2.7$. The censoring times are generated independently from an Exp(0.05) distribution.

The baseline cumulative hazard is examined for each of the unstratified Cox models under the following hazard rates,

$$\begin{aligned} &\alpha_1(t) = 0.2, \\ &\alpha_2(t) = 0.2t^{-0.5}, \qquad t > 0. \\ &\alpha_3(t) = 0.2e^{0.1t}, \end{aligned} \tag{45}$$

corresponding to constant, decreasing, and exponentially increasing hazard rates, respectively. The stratified model uses the two aforementoned covariates and stratum–specific hazard rates

$$\begin{aligned} &\alpha_1(t) = 0.2t, \\ &\alpha_2(t) = 0.2e^{0.1t}, \end{aligned} \qquad t > 0. \tag{46}$$

Survival times are generated using the inverse transform method (44), similar to the nonparametric comparison of Section 4.1.

4.3 Breslow–Cox vs. Breslow–GBM

The Breslow estimator of the bazeline cumulative hazard rate, for a Cox model, whose regression parameter vector $\boldsymbol{\beta}$ has been estimated via partial likelihood, is compared with the same estimator on a Cox model whose log relative risk function $f(\mathbf{x}_i)$ has been estimated with a GBM. The following two covariate structures are considered,

$$\begin{aligned} f_1(\mathbf{x}_i) &= 0.5x_{i1} - 0.7x_{i2}, \\ f_2(\mathbf{x}_i) &= \sin(\pi x_{i1}) + 0.5(2x_{i2} - 1), \end{aligned} \tag{47}$$

where $x_{i1} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $x_{i2} \overset{\text{iid}}{\sim} \text{Bin}(1, 0.5)$ are independent from each other. The two covariate structures are linear and nonlinear respectively.

The GBM is trained using survival times, censoring indicators, and covariate values. It is then fitted with a learning rate of $\eta = 0.01$ and a maximum of 3000 weak learners. Rather than fixing the tree depth and number of trees in advance, both hyperparameters are selected jointly by minimizing the ISE with respect to the true cumulative baseline hazard. This implies that the GBM is somewhat idealized, since we assume knowledge of the true cumulate baseline hazard when selecting tree depth and number of trees.

The tree depth is searched over the set $d \in \{1, 2, 3, 4\}$, and for each depth the number of trees is evaluated over a grid from 50 to 3000 in steps of 50. The combination of depth and tree count yielding the lowest ISE is then used to construct the final Breslow estimator. Thereafter, each covariate structure is examined under the following baseline hazard rates,

$$\begin{aligned} \alpha_1(t) &= 0.2, \\ \alpha_2(t) &= 0.2e^{0.1t}, \end{aligned} \qquad t > 0, \tag{48}$$

corresponding to constant and exponentially increasing hazard rates respectively. Censoring times are generated independently from an Exp(0.05) distribution and survival times are generated using the inverse transform method (44), similar to the nonparametric comparison of Section 4.1.

5 Results

Only figures featuring right-censoring are presented here, while the remaining figures are provided in the appendix because right–censoring is more common in practice.

5.1 Nelson–Aalen vs. Kaplan–Meier

The Nelson–Aalen and Kaplan–Meier estimators are compared under different hazard rates, both with and without right censoring.

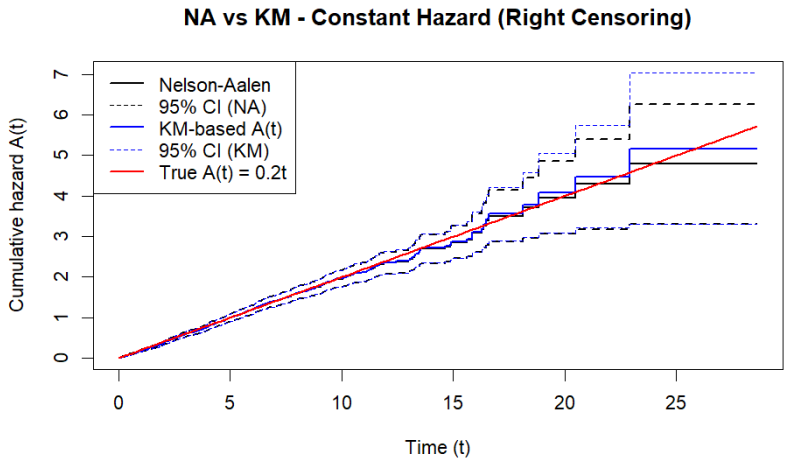

Figure 1: Nelson–Aalen (black) and Kaplan–Meier–based (blue) cumulative hazard estimates with 95% confidence intervals under a cumulative hazard $A(t) = 0.2t$ with 29.1% of the sample being right–censored, alongside the true cumulative hazard (red).

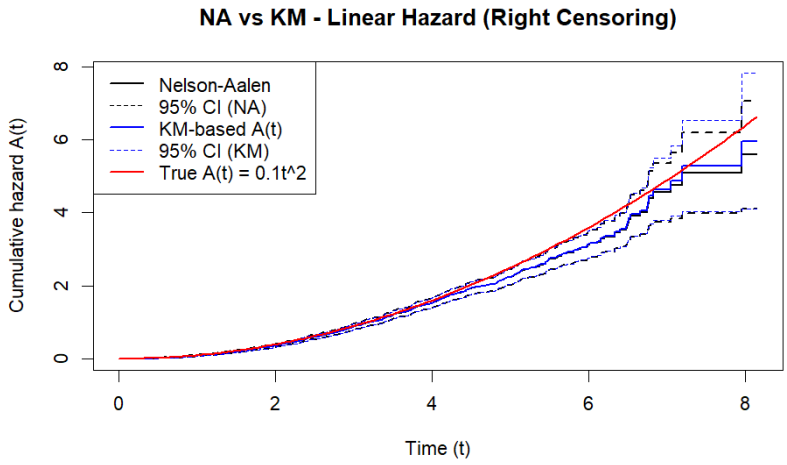

Figure 2: Nelson–Aalen (black) and Kaplan–Meier–based (blue) cumulative hazard estimates with 95% confidence intervals under a hazard $A(t) = 0.1t^2$ with 23.2% of the sample being right–censored, alongside the true cumulative hazard (red).

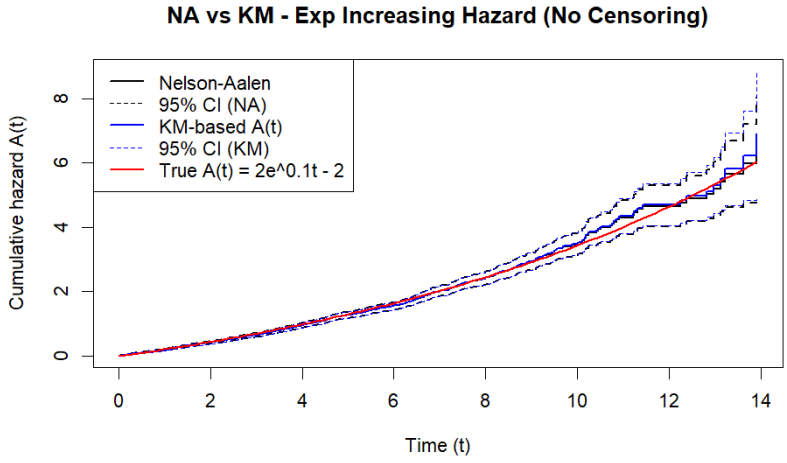

Figure 3: Nelson–Aalen (black) and Kaplan–Meier–based (blue) cumulative hazard estimates with 95% confidence intervals under a hazard $A(t) = 2e^{0.1t} - 2$, alongside the true cumulative hazard (red).

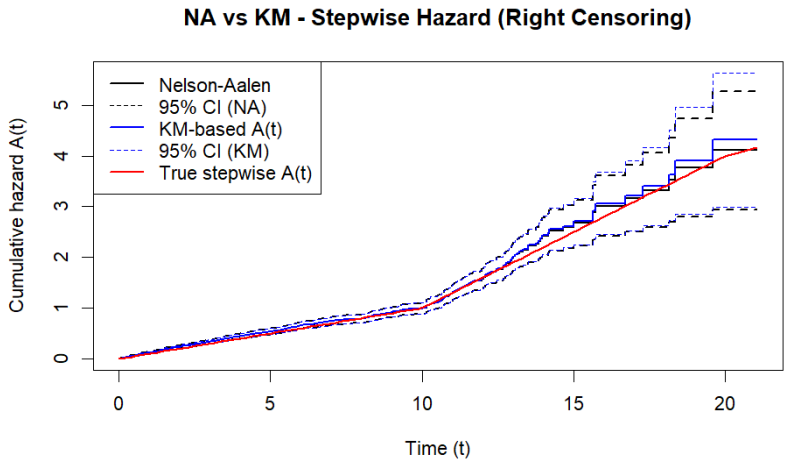

Figure 4: Nelson–Aalen (black) and Kaplan–Meier–based (blue) cumulative hazard estimates with 95% confidence intervals under a stepwise linear cumulative hazard $A(t)$ with 45.8% of the sample being right–censored, alongside the true cumulative hazard (red).

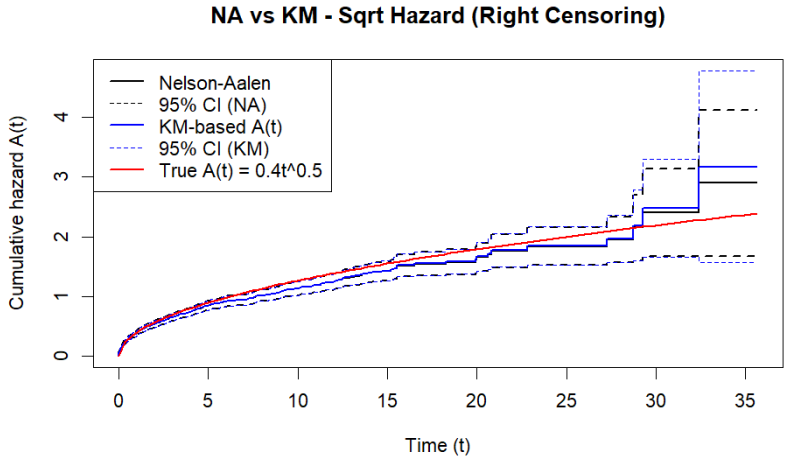

Figure 5: Nelson–Aalen (black) and Kaplan–Meier–based (blue) cumulative hazard estimates with 95% confidence intervals under a cumulative hazard $A(t) = 0.4t^{0.5}$ with 40.4% of the sample being right–censored, alongside the true cumulative hazard (red).

Table 2: Performance metrics Bias, ISE and SE for the Nelson–Aalen and Kaplan–Meier estimators without censoring. The best performing estimator for each metric is marked in bold.

Hazard $\alpha(t)$	**NA Bias**	**NA ISE**	**NA SE**	**KM Bias**	**KM ISE**	**KM SE**
Constant	**-0.01**	**3.37**	**0.48**	**-0.01**	5.00	1.22
Linear increasing	−0.05	0.65	0.53	**-0.04**	**0.50**	**0.09**
Exponential increasing	−0.018	**0.28**	**0.19**	**-0.01**	0.40	0.74
Stepwise	**0.06**	**3.22**	**0.95**	0.07	4.55	1.94
Decreasing	−0.02	**35.80**	**0.95**	**-0.01**	50.95	1.95

Table 3: Performance metrics Bias, ISE and SE for the Nelson–Aalen and Kaplan–Meier estimators with right–censoring. The best performing estimator for each metric is marked in bold.

Hazard $\alpha(t)$	**NA Bias**	**NA ISE**	**NA SE**	**KM Bias**	**KM ISE**	**KM SE**
Constant	**-0.02**	**0.49**	0.87	**-0.02**	2.59	**0.30**
Linear increasing	**-0.06**	0.43	1.10	**-0.06**	**0.32**	**0.45**
Exponential increasing	−0.04	**0.56**	**0.18**	**-0.03**	0.80	0.41
Stepwise	**0.03**	**0.38**	**0.00**	**0.03**	0.78	0.02
Decreasing	**-0.04**	**1.81**	**0.26**	**-0.04**	3.27	0.61

We see in Table 2 that the NA estimator records the lower ISE and SE for the constant, exponential increasing, stepwise, and decreasing hazard scenarios, while KM achieves the lower ISE and SE for the linear increasing hazard. Bias values are small and nearly identical between the two estimators across all hazard shapes. The decreasing hazard yields the largest ISE values overall, reaching 35.80 for NA and 50.95 for KM.

Table 3 shows that the results are more mixed under right–censoring. NA achieves the lower ISE for the constant, exponential increasing, stepwise, and decreasing hazards, while KM records the lower ISE for the linear increasing hazard. KM outperforms NA on SE for the constant and linear increasing hazards, whereas NA achieves the lower SE for the exponential increasing, stepwise, and decreasing cases. As in Table 2, bias values remain small and largely comparable between the two estimators.

Overall ISE values in Table 3 are generally smaller than their counterparts in Table 2 across most hazard shapes.

These results are consistent with the figures, where the two estimators overlap closely at early time points but diverge at later ones. The same is observed for their confidence intervals where KM's confidence intervals are broader than those of NA throughout.

5.2 Nelson–Aalen vs. Breslow–Cox

The Nelson–Aalen and Breslow estimators of the baseline cumulative hazard function are compared under different covariate structures, hazard rates, both with and without right censoring. Additionally, the estimators are compared in a stratified Cox model.

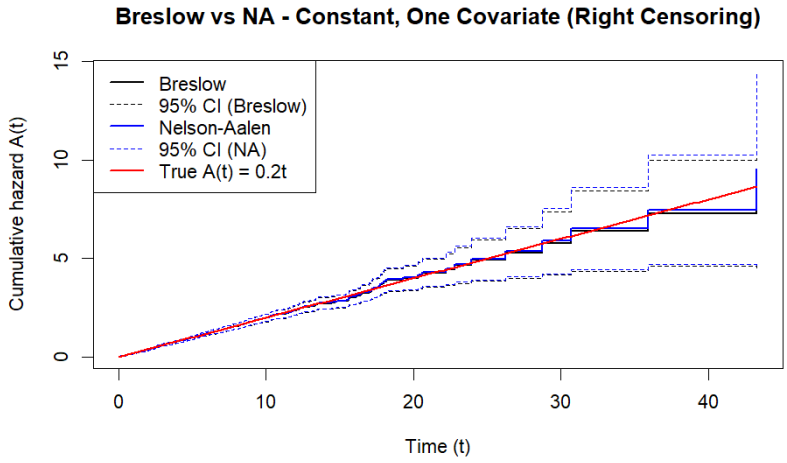

Figure 6: Nelson–Aalen and Breslow baseline cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$ with 19.4% of the sample being right–censored, alongside the true cumulative hazard (red). The singe covariate is standard normally distributed.

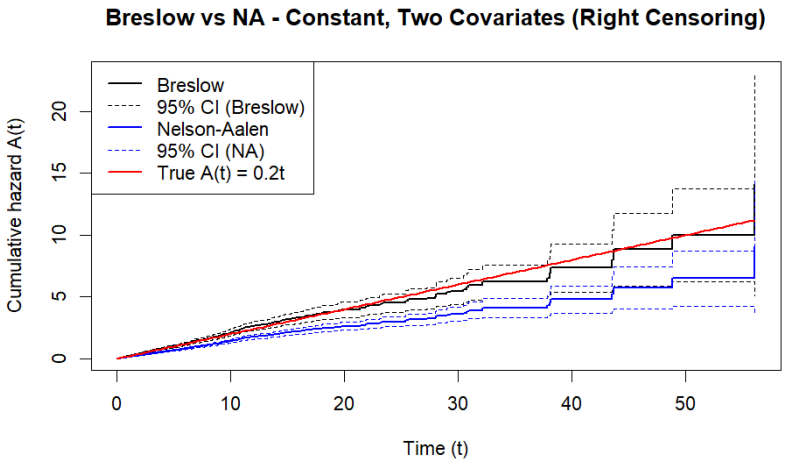

Figure 7: Nelson–Aalen and Breslow baseline cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$ with 19.4% of the sample being right–censored, alongside the true cumulative hazard (red). Two covariates are used one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

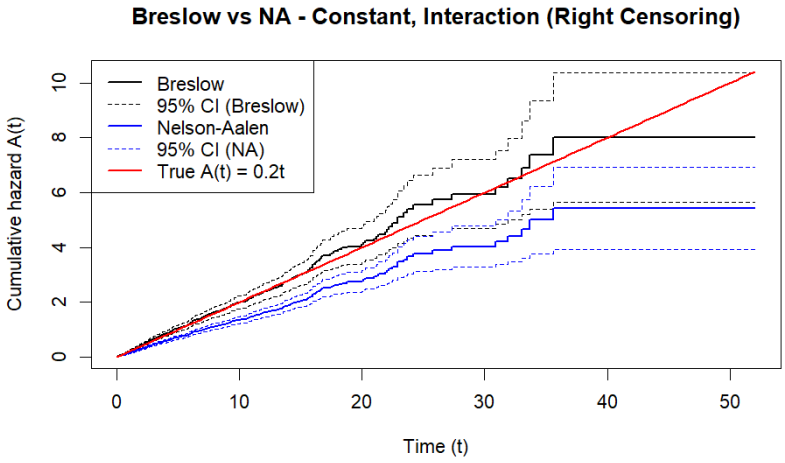

Figure 8: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$ with 27.9% of the sample being right–censored, alongside the true cumulative hazard (red). The same two covariates as in Figure 7 are used, together with a third covariate that corresponds to their interaction.

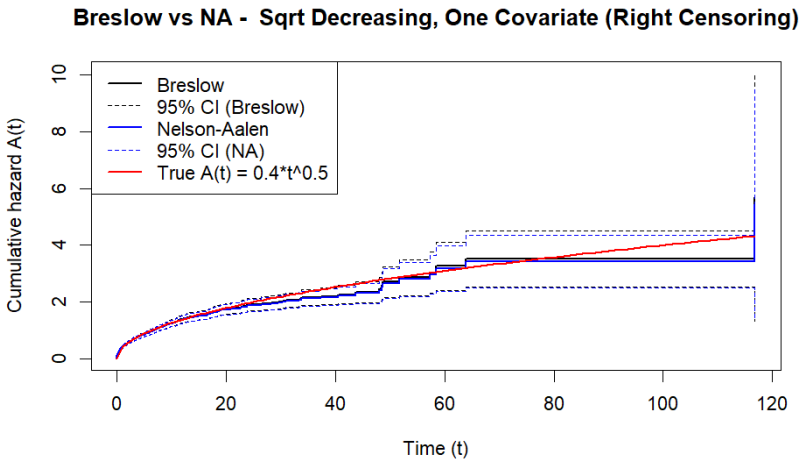

Figure 9: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.4t^{0.5}$ with 30.0% of the sample being right–censored, alongside the true cumulative hazard (red). The singe covariate is standard normally distributed.

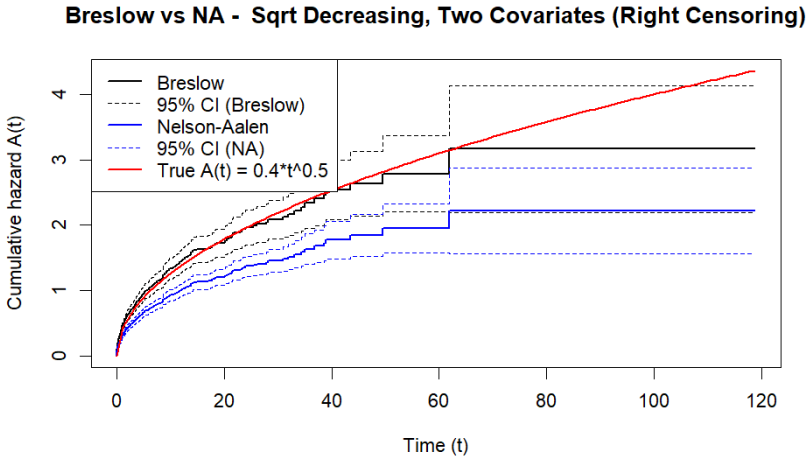

Figure 10: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.4t^{0.5}$ with 39.6% of the sample being right–censored, alongside the true cumulative hazard (red). Two covariates are used one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

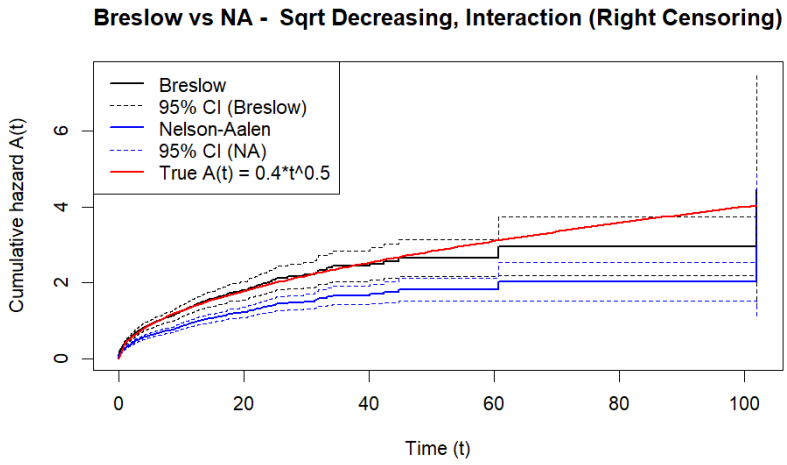

Figure 11: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.4t^{0.5}$ with 40.7% of the sample being right–censored, alongside the true cumulative hazard (red). The same two covariates as in Figure 10 are used, together with a third covariate that corresponds to their interaction.

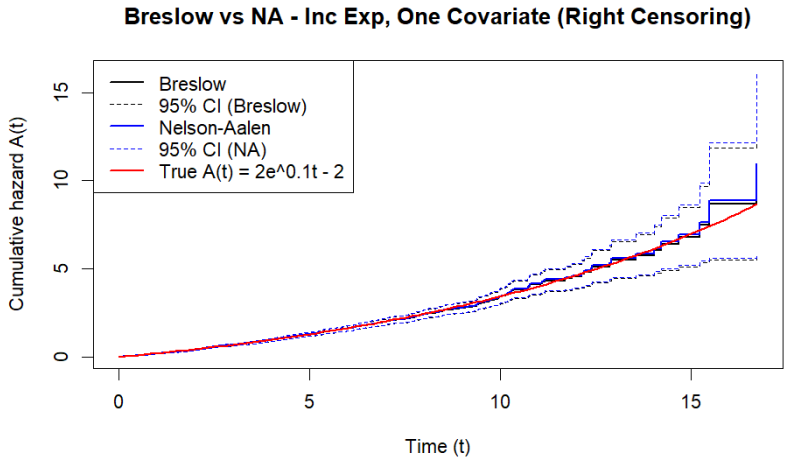

Figure 12: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$ with 15.1% of the sample being right–censored, alongside the true cumulative hazard (red). The singe covariate is standard normally distributed.

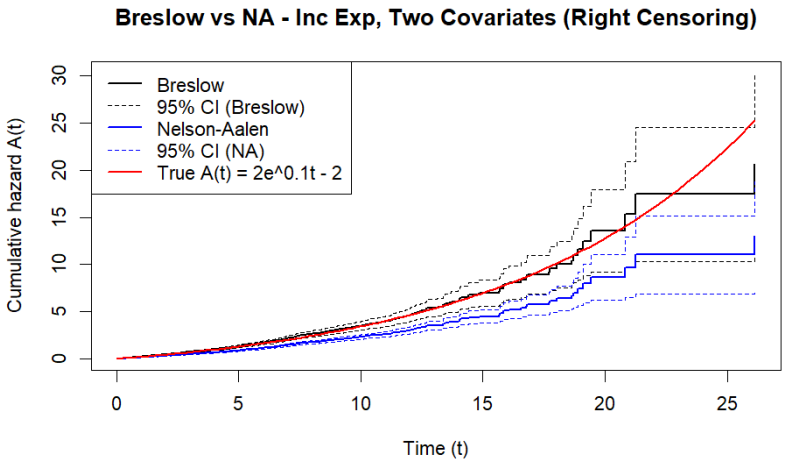

Figure 13: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$ with 20.9% of the sample being right–censored, alongside the true cumulative hazard (red). Two covariates are used one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

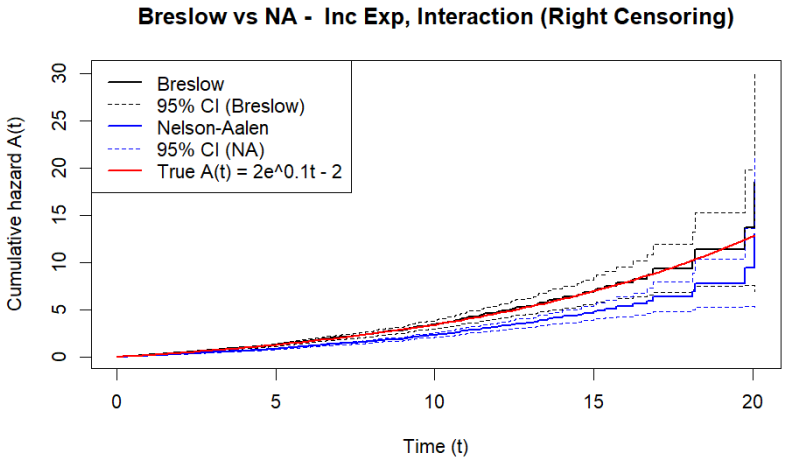

Figure 14: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$ with 20.5% of the sample being right–censored, alongside the true cumulative hazard (red). The same two covariates as in Figure 13 are used, together with a third covariate that corresponds to their interaction.

Table 4: Performance metrics Bias, ISE and SE for the Breslow and Nelson–Aalen estimators without censoring. The best performing estimator for each metric is marked in bold.

Baseline Hazard $\alpha_0(t)$	**B Bias**	**B ISE**	**B SE**	**NA Bias**	**NA ISE**	**NA SE**
Constant, 1 Cov	−0.01	**8.58**	**12.50**	**0.00**	11.80	14.54
Constant, 2 Cov	**0.12**	**161.10**	**4.14**	−0.59	3279.26	82.89
Constant, Interaction	**-0.16**	**667.66**	**39.98**	−0.56	2974.72	128.93
Decreasing, 1 Cov	**0.00**	1615.71	10.20	**0.00**	**961.91**	**7.30**
Decreasing, 2 Cov	**0.12**	**6161.40**	**4.14**	−0.58	162232.00	82.88
Decreasing, Interaction	**-0.04**	**3477.93**	**0.70**	−0.54	29801.76	19.34
Exp increasing, 1 Cov	−0.01	**1.72**	**12.50**	**0.00**	2.37	14.54
Exp increasing, 2 Cov	**0.12**	**19.88**	**4.14**	−0.59	325.16	82.88
Exp increasing, Interaction	**0.04**	**40.31**	**18.03**	$0 - 0.50$	61.29	94.84

Table 5: Performance metrics Bias, ISE and SE for the Breslow and Nelson–Aalen estimators with right–censoring. The best performing estimator for each metric is marked in bold.

Baseline Hazard $\alpha_0(t)$	**B Bias**	**B ISE**	**B SE**	**NA Bias**	**NA ISE**	**NA SE**
Constant, 1 Cov	−0.01	**0.62**	**0.38**	**0.00**	1.83	0.80
Constant, 2 Cov	**0.02**	**6.51**	**8.26**	−0.38	293.10	4.61
Constant, Interaction	**-0.08**	**8.62**	1.43	−0.35	67.93	**0.78**
Decreasing, 1 Cov	0.03	**3.85**	**0.00**	**0.02**	4.66	**0.00**
Decreasing, 2 Cov	**0.02**	**16.11**	**1.21**	−0.28	168.09	4.88
Decreasing, Interaction	**0.03**	**6.40**	**0.23**	−0.26	89.63	2.88
Exp increasing, 1 Cov	−0.01	**2.12**	**4.10**	**0.00**	2.90	5.25
Exp increasing, 2 Cov	**0.04**	**39.76**	**21.89**	−0.44	154.63	150.91
Constant, Interaction	**0.02**	**2.84**	31.97	−0.38	46.54	**0.02**

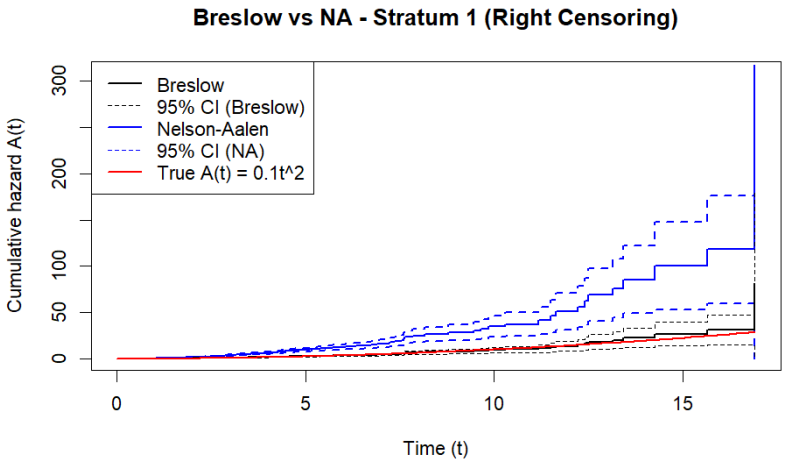

Figure 15: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_{10}(t) = 0.1t^2$ for stratum 1 with 10.6% of the sample being right–censored, alongside the true cumulative hazard (red). Two covariates are used, of which one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

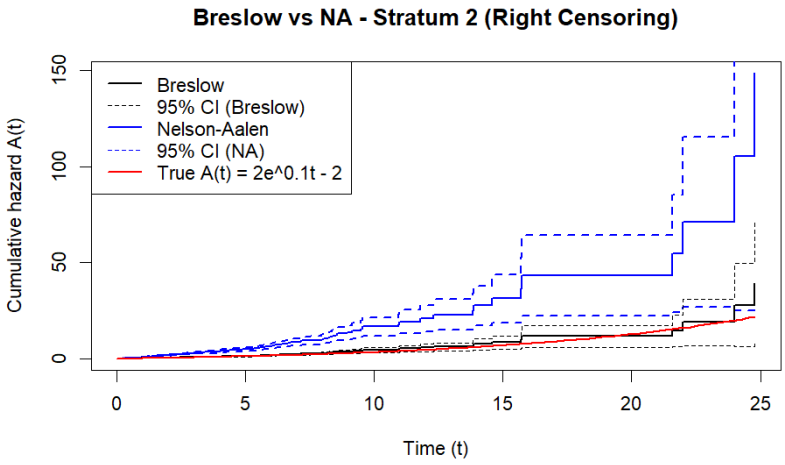

Figure 16: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_{20}(t) = 2e^{0.1t} - 2$ for stratum 2 with 10.6% of the sample being right–censored, alongside the true cumulative hazard (red). Two covariates are used, of which one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

Table 6: Performance metrics Bias, ISE and SE for the Breslow and Nelson–Aalen estimators for two strata with two covariates without censoring. The best performing estimator for each metric is marked in bold.

Stratum Baseline Hazard $\alpha_{s0}(t)$	**B Bias**	**B ISE**	**B SE**	**NA Bias**	**NA ISE**	**NA SE**
Linear increasing	**0.04**	**5977.20**	**3011.00**	5.60	422328.70	348886.80
Exp increasing	**0.43**	**2864.52**	**1601.78**	4.30	123880.85	54624.39

Table 7: Performance metrics Bias, ISE and SE for the Breslow and Nelson–Aalen estimators for two strata with two covariates with right–censoring. The best performing estimator for each metric is marked in bold.

Stratum Basline Hazard $\alpha_{s0}(t)$	**B Bias**	**B ISE**	**B SE**	**NA Bias**	**NA ISE**	**NA SE**
Linear increasing	**0.17**	**140.11**	**2780.12**	3.32	30483.76	78559.22
Exp increasing	**0.16**	**167.08**	**308.85**	2.44	21989.22	16117.88

In Table 4, the Breslow estimator achieves the lower ISE and SE across nearly all scenarios without censoring, with the exception of the decreasing baseline hazard with one covariate. Bias values are small and comparable between the two estimators in the one–covariate scenarios, while Nelson–Aalen exhibits notably larger negative bias in the two–covariate and interaction settings. The largest ISE values are observed under the decreasing baseline hazard, particularly for

the two–covariate and interaction cases.

We see in Table 5 that the Breslow estimator continues to achieve the lower ISE across all scenarios under right–censoring, and the lower SE in the majority of cases. Nelson–Aalen records a lower SE only in the constant interaction and decreasing one-covariate scenarios. As seen in Table 4, Nelson–Aalen exhibits larger negative bias in the two–covariate and interaction settings, while bias values remain small and comparable in the one–covariate scenarios. Overall, ISE values are substantially reduced compared to Table 4 across all hazard shapes and covariate structures.

In Table 6, the Breslow estimator outperforms Nelson–Aalen on all metrics for both stratum baseline hazard shapes in the stratified setting without censoring. Nelson–Aalen exhibits substantially larger bias, ISE, and SE, with ISE values several orders of magnitude larger than those of the Breslow estimator.
We see these results again in Table 7, where the Breslow estimator achieves lower bias, ISE, and SE than Nelson–Aalen for both baseline hazard shapes under right-censoring. As in Table 6, Nelson–Aalen records substantially larger values across all metrics, though the ISE values for both estimators are considerably reduced compared to the uncensored stratified setting.

The figures show that the estimators and their respective 95% confidence intervals behave very similarly in settings with one covariate. In non–stratified scenarios with more than one covariate, Nelson–Aalen severely underestimates, which is consistent with the results in Tables 4 and 5. In stratified scenarios with several covariates, Nelson–Aalen instead overestimates and explodes at the last time point, consistent with the substantially larger ISE and SE values seen in Tables 6 and 7.

5.3 Breslow–Cox vs. Breslow–GBM

The Breslow estimator with and without the aid of GBMs are compared under different log relative risk functions, hazard rates, both with and without right censoring.

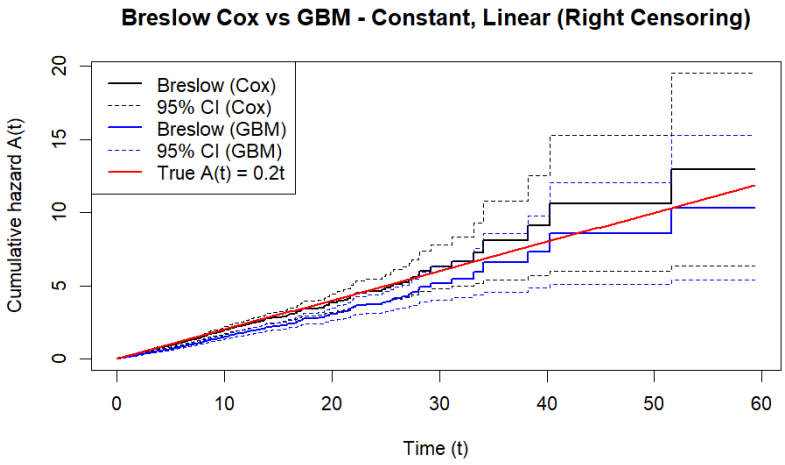

Figure 17: The Breslow cumulative hazard estimates with and without the aid of a GBM estimating $f(\mathbf{x}_i) = 0.5x_{i1} - 0.7x_{i2}$. Their respective 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$ with 30.2% of the sample being right–censored, alongside the true cumulative hazard (red).

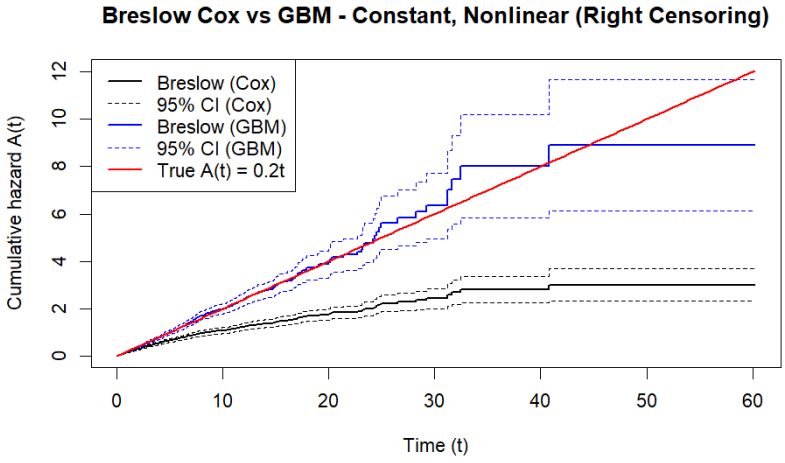

Figure 18: The Breslow cumulative hazard estimates with and without the aid of a GBM estimating $f(\mathbf{x}_i) = \sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$. Their respective 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$ with 24.6% of the sample being right–censored, alongside the true cumulative hazard (red).

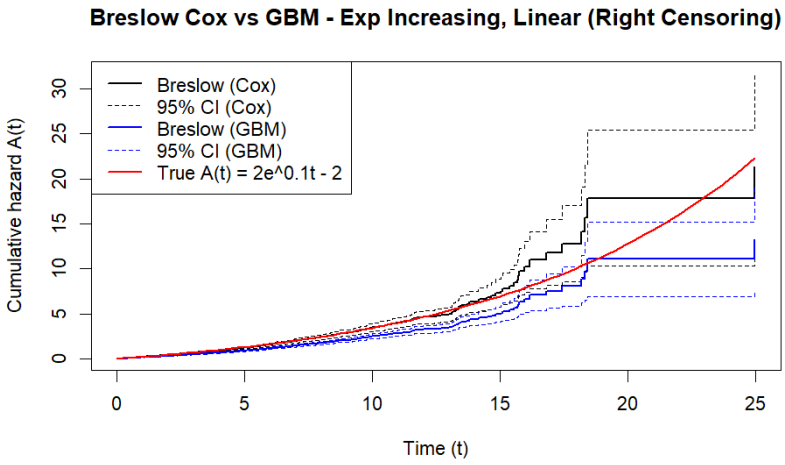

Figure 19: The Breslow cumulative hazard estimates with and without the aid of a GBM estimating $f(\mathbf{x}_i) = 0.5x_{i1} - 0.7x_{i2}$. Their respective 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$ with 22.4% of the sample being right–censored, alongside the true cumulative hazard (red).

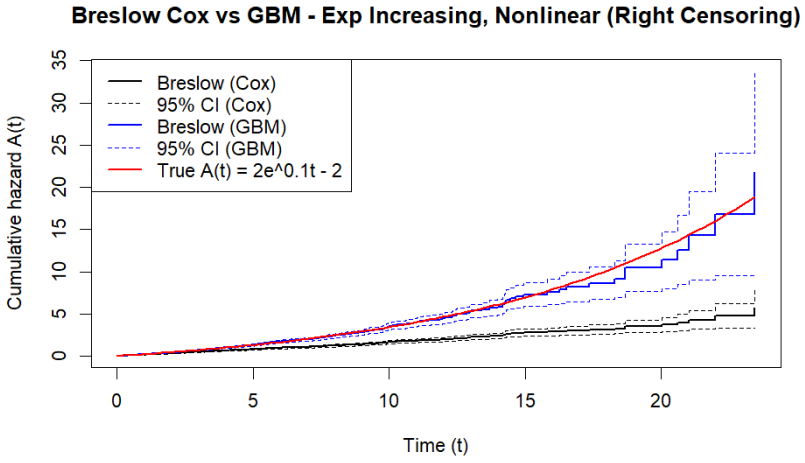

Figure 20: The Breslow cumulative hazard estimates with and without the aid of a GBM estimating $f(\mathbf{x}_i) = \sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$. Their respective 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$ with 18.7% of the sample being right–censored, alongside the true cumulative hazard (red).

Table 8: The optimal parameters of GBMs estimating $\mathbf{f}(\mathbf{x}_i)$ with and without right–censoring with respect to the ISE regarding the true baseline cumulative hazard rate $A_0(t)$.

Baseline Hazard $\alpha_0(t)$	**Censoring**	$\mathbf{f}(\mathbf{x}_i)$	**Tree Depth**	**Number of Weak Learners**
Constant	No	$0.5x_{i1} - 0.7x_{i2}$	2	1900
Constant	Yes	$0.5x_{i1} - 0.7x_{i2}$	3	1200
Exp increasing	No	$0.5x_{i1} - 0.7x_{i2}$	3	1000
Exp increasing	Yes	$0.5x_{i1} - 0.7x_{i2}$	4	200
Constant	No	$\sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$	1	2050
Constant	Yes	$\sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$	1	3000
Exp increasing	No	$\sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$	1	2050
Exp increasing	Yes	$\sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$	1	2300

In Table 8, the optimal tree depth ranges from 1 to 4 and the number of weak learners ranges from 200 to 3000 across the different scenarios. For the linear log relative risk function $f_1(\mathbf{x}_i)$, the optimal tree depth tends to be larger and the number of weak learners smaller compared to the nonlinear function $f_2(\mathbf{x}_i)$, where a tree depth of 1 is optimal across all scenarios.

Table 9: Performance metrics Bias, ISE and SE for the Breslow estimator with and without the assistance of a GBM estimating the log relative risk function $f_1(\mathbf{x}_i) = 0.5x_{i1} - 0.7x_{i2}$, without censoring. The best performing estimator for each metric is marked in bold.

Basline Hazard $\alpha_0(t)$	**No GBM Bias**	**No GBM ISE**	**No GBM SE**	**GBM Bias**	**GBM ISE**	**GBM SE**
Constant	**0.06**	817.35	**1.26**	−0.36	**219.66**	9.05
Exp increasing	**0.06**	102.96	**1.26**	−0.36	**27.06**	11.80

Table 10: Performance metrics Bias, ISE and SE for the Breslow estimator with and without the assistance of a GBM estimating the log relative risk function $f_1(\mathbf{x}_i) = 0.5x_{i1} - 0.7x_{i2}$, with right–censoring. The best performing estimator for each metric is marked in bold.

Basline Hazard $\alpha_0(t)$	**No GBM Bias**	**No GBM ISE**	**No GBM SE**	**GBM Bias**	**GBM ISE**	**GBM SE**
Constant	**-0.04**	141.06	**1.17**	−0.29	**21.06**	2.34
Exp increasing	**0.00**	373.44	**0.82**	−0.36	**21.45**	82.11

Table 9 shows that the GBM–assisted Breslow estimator achieves the lower ISE for both baseline hazard shapes without censoring under the linear log relative risk function $f_1(\mathbf{x}_i)$, while the Breslow estimator without GBM has the lower bias and SE. The reduction in ISE from using the GBM is substantial, decreasing from 817.35 to 219.66 for the constant hazard and from 102.96 to 27.06 for the exponentially increasing hazard.

In Table 10, the GBM-assisted Breslow estimator again achieves the lower ISE under right-censoring for both baseline hazard shapes, while the Breslow estimator without GBM has the lower bias and SE. The ISE reductions are again considerable, and overall ISE values are substantially lower than those in Table

9 for both estimators.

Table 11: Performance metrics Bias, ISE and SE for the Breslow estimator with and without the assistance of a GBM estimating the log relative risk function $f_2(\mathbf{x}_i) = \sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$, without censoring. The best performing estimator for each metric is marked in bold.

Basline Hazard $\alpha_0(t)$	**No GBM Bias**	**No GBM ISE**	**No GBM SE**	**GBM Bias**	**GBM ISE**	**GBM SE**
Constant	−0.69	7147.71	238.88	**0.00**	**12.59**	**0.04**
Exp increasing	−0.69	816.62	238.88	**0.00**	**1.45**	**0.04**

Table 12: Performance metrics Bias, ISE and SE for the Breslow estimator with and without the assistance of a GBM estimating the log relative risk function $f_2(\mathbf{x}_i) = \sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$, with right–censoring. The best performing estimator for each metric is marked in bold.

Basline Hazard $\alpha_0(t)$	**No GBM Bias**	**No GBM ISE**	**No GBM SE**	**GBM Bias**	**GBM ISE**	**GBM SE**
Constant	−0.43	1123.59	81.87	**0.00**	**47.94**	**9.96**
Exp increasing	−0.53	607.83	173.87	**-0.01**	**4.52**	**8.36**

We see in Table 11 that the GBM-assisted Breslow estimator outperforms the estimator without GBM on all metrics for both baseline hazard shapes without censoring under the nonlinear log relative risk function $f_2(\mathbf{x}_i)$. The improvement is considerably larger than in the linear case, with ISE reducing from 7147.71 to 12.59 for the constant hazard and from 816.62 to 1.45 for the exponentially increasing hazard. The bias of the estimator without GBM is also substantially larger at −0.69 for both hazard shapes, compared to 0.00 for the GBM-assisted estimator.

In Table 12, the GBM-assisted Breslow estimator again outperforms the estimator without GBM on all metrics under right-censoring for both baseline hazard shapes. As in Table 11, the estimator without GBM exhibits large negative bias and substantially higher ISE and SE, while the GBM-assisted estimator achieves near-zero bias and considerably lower ISE and SE across both hazard rates.

6 Discussion

6.1 Analysis of Obtained Results

The reason that the Nelson–Aalen and Kaplan–Meier estimators of the cumulative hazard rate overlap at early time points and Kaplan–Meier has larger estimate of the cumulative hazard rate than Nelson–Aalen as time increases is that the two estimators are related by the approximation

$$-\ln\left(1-\frac{1}{Y(T_j)}\right)\approx\frac{1}{Y(T_j)}, \tag{49}$$

which follows from the Taylor expansion $-\ln(1-x)=x+\frac{x^2}{2}+\frac{x^3}{3}+\cdots$ evaluated at $x=\frac{1}{Y(T_j)}$. At early time points the risk set $Y(T_j)$ is large, so $\frac{1}{Y(T_j)}$ is small and the higher–order terms are negligible, meaning that the two estimators produce nearly identical increments at each event time and their cumulative sums coincide. As time progresses, the risk set shrinks due to previous events and censoring, so that the individual jumps $\frac{1}{Y(T_j)}$ grow larger and the discarded higher–order terms become non–negligible. Since $-\ln\left(1-\frac{1}{Y(T_j)}\right)>\frac{1}{Y(T_j)}$ for all $Y(T_j)>0$, each increment of the Kaplan–Meier transformation (12) strictly exceeds the corresponding increment of the Nelson–Aalen estimator (8) and this gap compounds over time, causing the Kaplan–Meier curve to drift above Nelson–Aalen toward the right tail.

The confidence intervals get wider with time because the variance estimator (9) accumulates a new term $\frac{1}{Y(T_j)^2}$ at each event time and as the risk set $Y(T_j)$ shrinks due to previous events and censoring, these terms grow larger. Consequently $\hat{\sigma}$ increases monotonically with t, and since the Nelson–Aalen confidence interval half–width is $z_{1-\alpha/2}\hat{\sigma}$, the bounds continuously get wider. The Breslow estimator carries two sources of uncertainty, the estimation of $\hat{\boldsymbol{\beta}}$ via partial likelihood and the estimation of the baseline hazard itself. Both of these sources of uncertainty grow with time as the risk set shrinks, contributing to similarly widening bounds.

In some figures such as Figure 9, the lower confidence bound of an estimator decreases at late time points. For the Nelson–Aalen estimator, this occurs when so few individuals remain at risk that $\hat{\sigma}$ grows faster than $\hat{A}(t)$ accumulates, causing the lower bound $\hat{A}(t)-z_{1-\alpha/2}\hat{\sigma}$ to turn negative and decrease. For the Kaplan–Meier estimator this instability is further amplified by the transformation (13), since the lower bound is $-\ln(\hat{S}(t)+z_{1\alpha/2}\hat{\sigma})$ and as $\hat{A}(t)$ and $\hat{\sigma}$ both grow large at late time points, their sum increases and its negative logarithm decreases. For the Breslow estimator, the lower bound decreases for the same reason as for Nelson–Aalen estimator.

The comparably larger ISE values observed under the hazard rate $\alpha(t)=0.2t^{-0.5}$ are a consequence of the longer time scale over which the estimator

is evaluated. Since the hazard rate is decreasing, risk is highest early and diminishes over time, which causes individuals to survive longer on average. This pushes the last observed survival time T to larger values and since the ISE integrates squared errors over $[0, T]$, a longer time range directly inflates the ISE even if the point–wise estimation error is comparable to other hazard shapes.

Across all comparisons, no single estimator dominates universally, rather the relative performance depends on the covariate structure, hazard shape and whether the primary interest lies in overall accuracy or tail behavior.

For the Nelson–Aalen and Kaplan–Meier comparison, neither estimator consistently outperforms the other. Nelson–Aalen achieves lower ISE in most scenarios, particularly under constant, stepwise, and decreasing hazards, while Kaplan–Meier achieves lower ISE under the linearly increasing hazard. Since bias is nearly identical throughout, the difference is most apparent in the right tail, where Kaplan–Meier's broader confidence intervals reflect the larger margin of error. Nelson–Aalen is preferable, especially when tail accuracy is important, while Kaplan–Meier is a natural choice when the survival function itself is of primary interest.

For the Nelson–Aalen and Breslow comparison, the Breslow estimator dominates as soon as more than one covariate is present. Both estimators use the weighted risk set $\sum_{l=1}^{n} Y_l(T_j) r(\boldsymbol{\beta}, \mathbf{x}_l(T_j))$ in their denominators, with Nelson–Aalen using the true $\boldsymbol{\beta}$ and Breslow substituting $\hat{\boldsymbol{\beta}}$ estimated via partial likelihood. Counterintuitively, using the true $\boldsymbol{\beta}$ performs worse. This is can be explained with the fact that $\hat{\boldsymbol{\beta}}$ is estimated from the same data and adapts to the realized covariate values and effectively compensating for sampling variability, whereas the true $\boldsymbol{\beta}$ makes no such adaptation. The advantage of Breslow holds across all hazard shapes and covariate structures, and is even more pronounced in the stratified setting.

For the Breslow–Cox and Breslow–GBM comparison, GBM assistance reduces ISE substantially in all scenarios. Under the linear covariate structure $f_1(\mathbf{x}_i)$, the improvement is meaningful but moderate, since Cox's partial likelihood already estimates a linear log relative risk consistently with low bias and SE. However it achieves higher ISE values. Under the nonlinear structure $f_2(\mathbf{x}_i)$, the Cox model is unsuitable and produces large negative bias and higher SE, while the GBM captures the nonlinear structure and reduces ISE remarkably. The GBM-assisted Breslow estimator never performs much worse than the Breslow-Cox estimator, and therefore seems preferable, especially when the log relative risk function $f(\mathbf{x}_i)$ is assumed to be nonlinear. Furthermore, it should be noted that the GBM-assisted Breslow estimator operates under the idealized assumption that the true baseline cumulative hazard function $A_0(t)$ is known, as described in Section 5.3.

6.2 Possible Extensions

The simulation study could be extended in several directions. First, the analyses were conducted with a fixed sample size of $n = 1000$, and repeating them with smaller or larger samples would give insights on how the estimators behave as the amount of available data varies. Second, only a limited selection of hazard rate functions was considered and other shapes such as bathtub or hump–shaped hazard rates, which are common in reliability and medical contexts, could reveal different patterns of estimator performance. Third, the covariates considered were all constant, whereas many real applications involve time–dependent covariates such as a patient's measured blood pressure at each follow–up visit. Extending the Cox and GBM frameworks to time–dependent covariates would make the comparisons more realistic. Fourth, the SE metric was evaluated at the last observed survival time t_k, since estimation uncertainty is greatest there due to the smallest remaining risk set. An alternative would be to report the maximum SE, $\max_i \left(\hat{A}(t_i) - A(t_i)\right)^2$, over all observed time points, which would identify the worst–case point–wise deviation regardless of where it occurs along the time axis. Finally, all censoring in this study was random and independent of the survival time, generated from a fixed exponential distribution. In practice, censoring is often more structured, for instance in a clinical trial where all individuals are enrolled over a recruitment period and the study ends at a fixed calendar date, so that individuals enrolled later have shorter potential follow–up times than those enrolled earlier. It would perhaps be a more realistic approach.

In the Breslow–GBM comparison, the optimal tree depth and number of weak learners were selected by minimizing the ISE with respect to the true cumulative baseline hazard $A_0(t)$. This was possible because the true hazard rate was known in the simulation setting, making ISE a natural and direct criterion for hyperparameter selection. In practice however, the true $A_0(t)$ is unknown and alternative selection methods must be used. A candidate is K-fold cross–validation where in the Cox–GBM setting this validation error would be the negative partial log–likelihood (33) evaluated on the held–out fold. Comparing these strategies against the ISE criterion used here would be a natural extension of the study, as it would assess how much performance is lost when the true hazard is unknown and data–driven hyperparameter tuning must be used instead.

7 References

Aalen, Odd O., Borgan, Ørnulf & Gjessing, Håkon K. *Survival and Event History Analysis*. 1st ed. (Springer, 2008).

Held, Leonhard, & Bové, Daniel Sabanés. *Likelihood and Bayesian Inference: With Applications in Biology and Medicine*. 2nd ed. (Springer, 2020).

GeeksforGeeks. *Decision Tree in Machine Learning*. https://www.geeksforgeeks.org/machine-learning/decision-tree/ (accessed 10 March 2026).

Yuksel, Mutlu & Aydede, Yigit. *Causal Inference and Machine Learning: In Economics, Social, and Health Sciences*. https://www.causalmlbook.com/ (accessed 12 March 2026).

Natekin, Alexey & Knoll, Alois. *Gradient Boosting Machines, a Tutorial*. *Frontiers in Neurorobotics*, vol. 7, art. 21 (2013). https://pmc.ncbi.nlm.nih.gov/articles/PMC3885826/ (accessed 12 March 2026).

Casella, George & Berger, Roger L. *Statistical Inference*. 2nd ed. (Duxbury Press, 2002).

Emergent Mind. Integrated Mean Squared Error. https://www.emergentmind.com/topics/integrated-mean-squared-error (accessed 25 March 2026).

8 Appendix

8.1 Nelson–Aalen vs. Kaplan–Meier

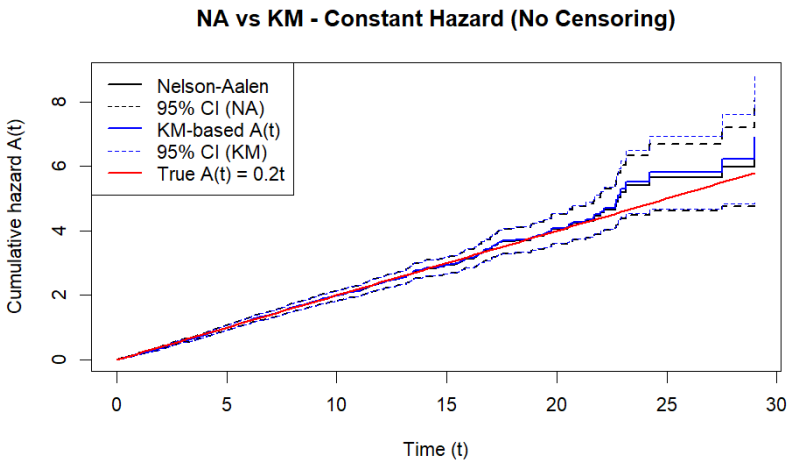

Figure 21: Nelson–Aalen (black) and Kaplan–Meier–based (blue) cumulative hazard estimates with 95% confidence intervals under a cumulative hazard $A(t) = 0.2t$, alongside the true cumulative hazard (red).

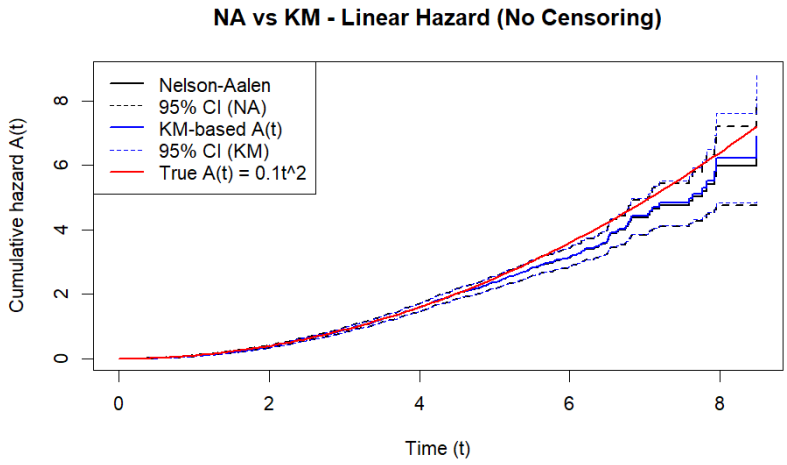

Figure 22: Nelson–Aalen and Kaplan–Meier–based cumulative hazard estimates with 95% confidence intervals under a cumulative hazard $A(t) = 0.2t$, alongside the true cumulative hazard (red).

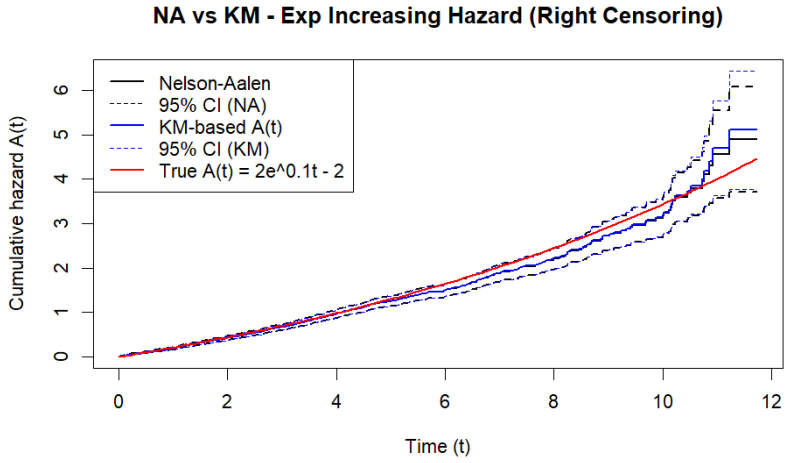

Figure 23: Nelson–Aalen (black) and Kaplan–Meier–based (blue) cumulative hazard estimates with 95% confidence intervals under a hazard $A(t) = 2e^{0.1t} - 2$ with 28.0% of the sample being right–censored, alongside the true cumulative hazard (red).

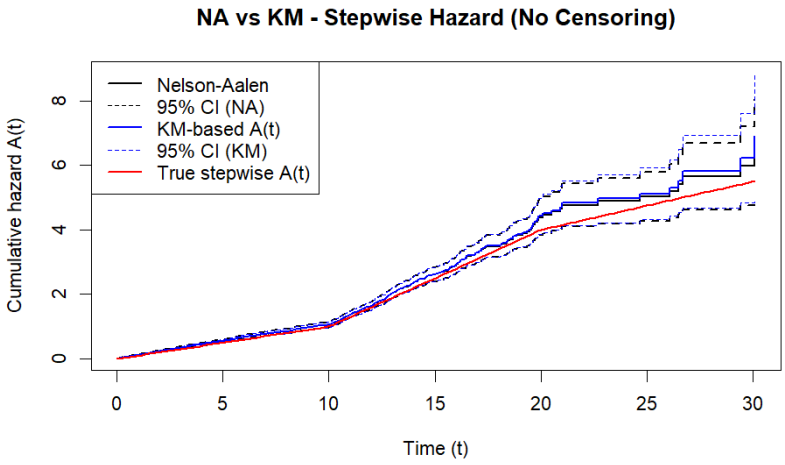

Figure 24: Nelson–Aalen (black) and Kaplan–Meier–based (blue) cumulative hazard estimates with 95% confidence intervals under a stepwise cumulative hazard $A(t)$, alongside the true cumulative hazard (red).

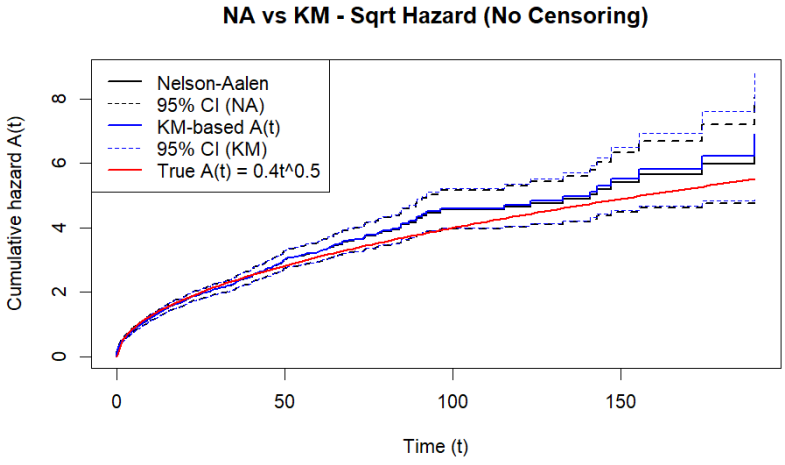

Figure 25: Nelson–Aalen (black) and Kaplan–Meier–based (blue) cumulative hazard estimates with 95% confidence intervals under a cumulative hazard $A(t) = 0.4t^{0.5}$, alongside the true cumulative hazard (red).

8.2 Nelson–Aalen vs. Breslow–Cox

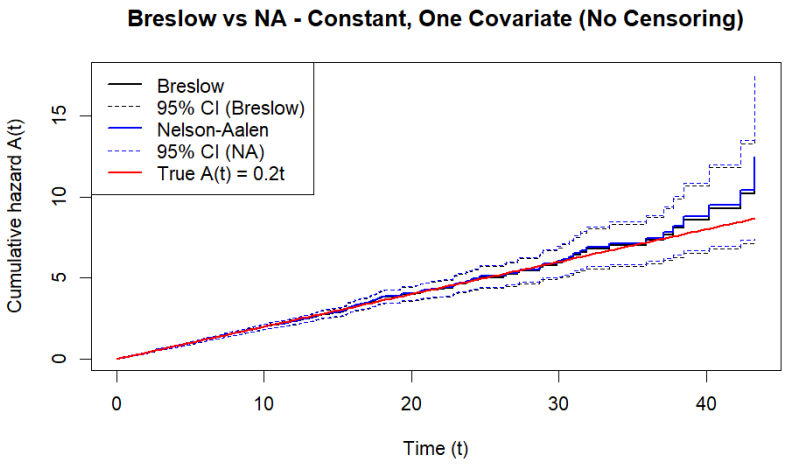

Figure 26: Nelson–Aalen and Breslow baseline cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$, alongside the true cumulative hazard (red). The singe covariate is standard normally distributed.

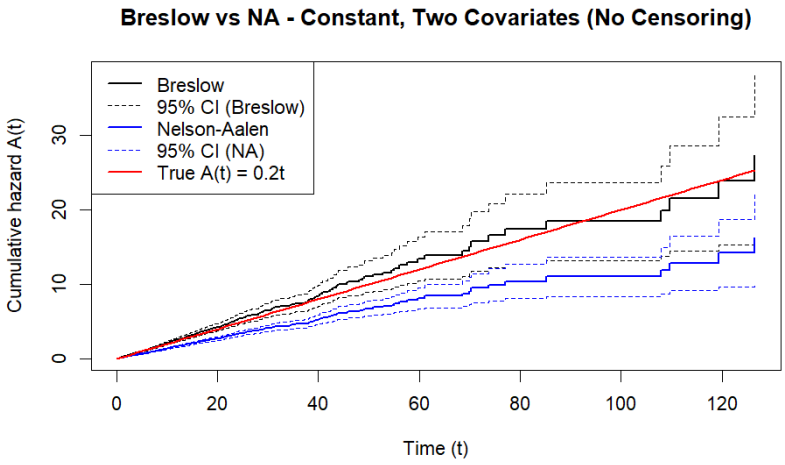

Figure 27: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$, alongside the true cumulative hazard (red). Two covariates are used one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

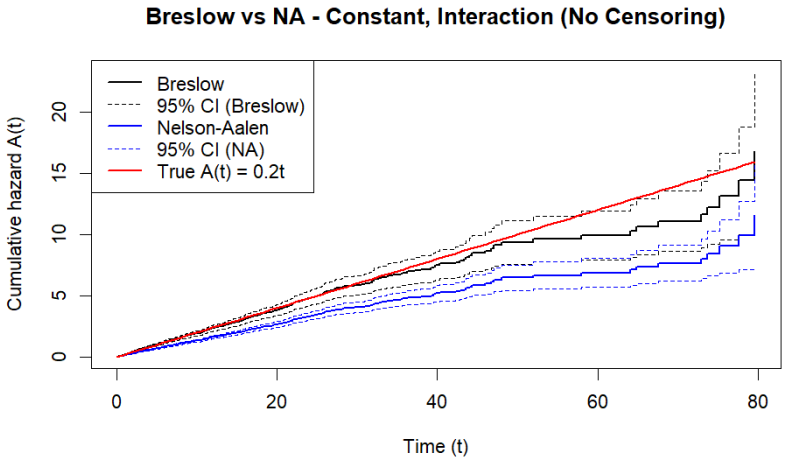

Figure 28: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$, alongside the true cumulative hazard (red). The same two covariates as in Figure 27 are used, together with a third covariate that corresponds to their interaction.

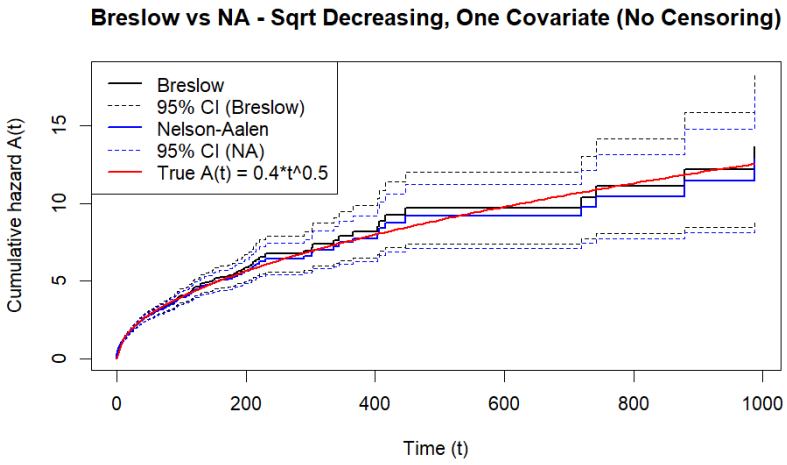

Figure 29: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.4t^{0.5}$, alongside the true cumulative hazard (red). The singe covariate is standard normally distributed.

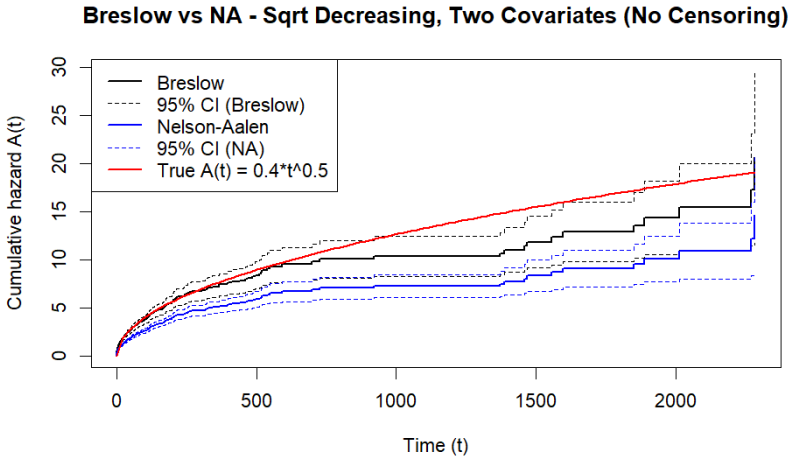

Figure 30: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.4t^{0.5}$, alongside the true cumulative hazard (red). Two covariates are used one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

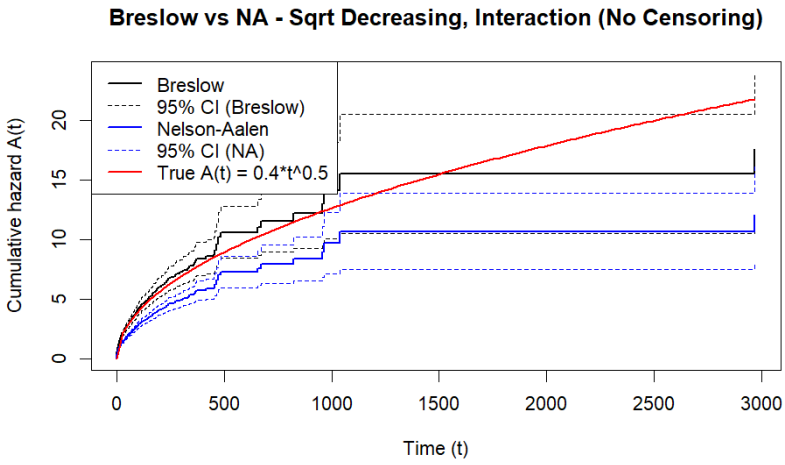

Figure 31: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.4t^{0.5}$, alongside the true cumulative hazard (red). The same two covariates as in Figure 30 are used, together with a third covariate that corresponds to their interaction.

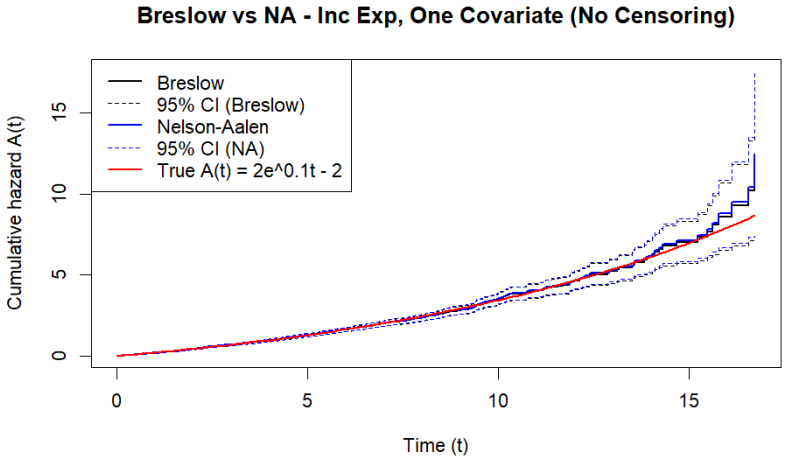

Figure 32: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$, alongside the true cumulative hazard (red). The singe covariate is standard normally distributed.

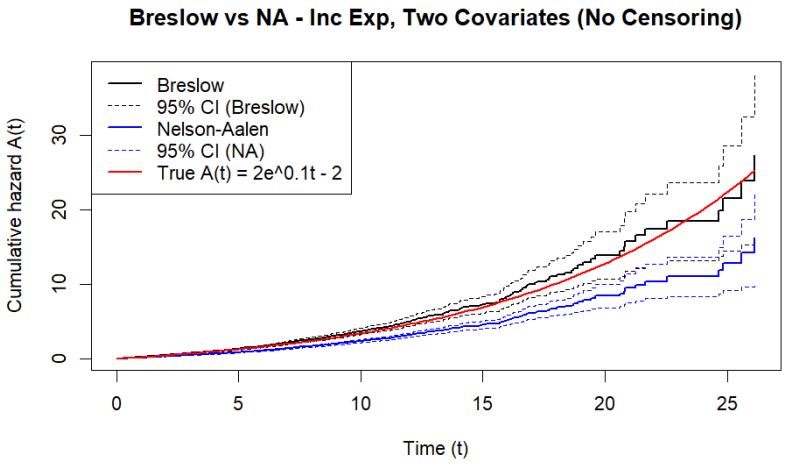

Figure 33: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$, alongside the true cumulative hazard (red). Two covariates are used one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

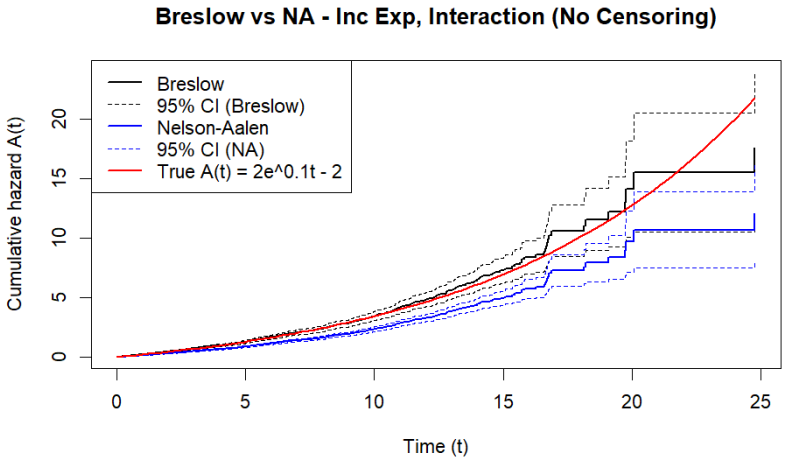

Figure 34: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$, alongside the true cumulative hazard (red). The same two covariates as in Figure 33 are used, together with a third covariate that corresponds to their interaction.

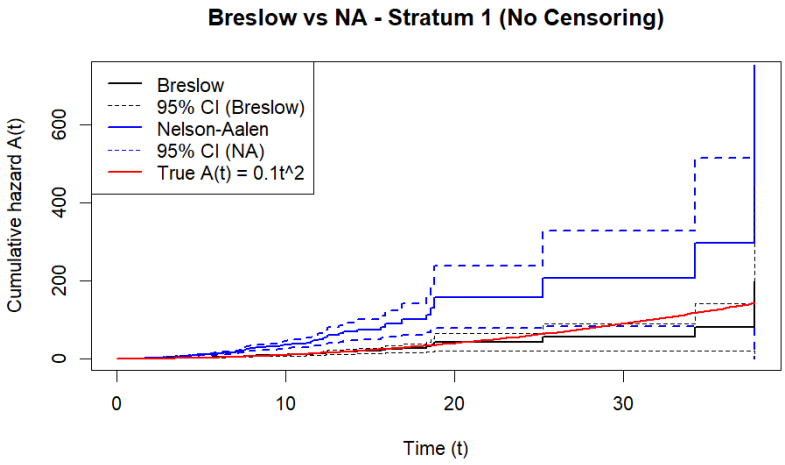

Figure 35: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_{10}(t) = 0.1t^2$ for stratum 1, alongside the true cumulative hazard (red). Two covariates are used, of which one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

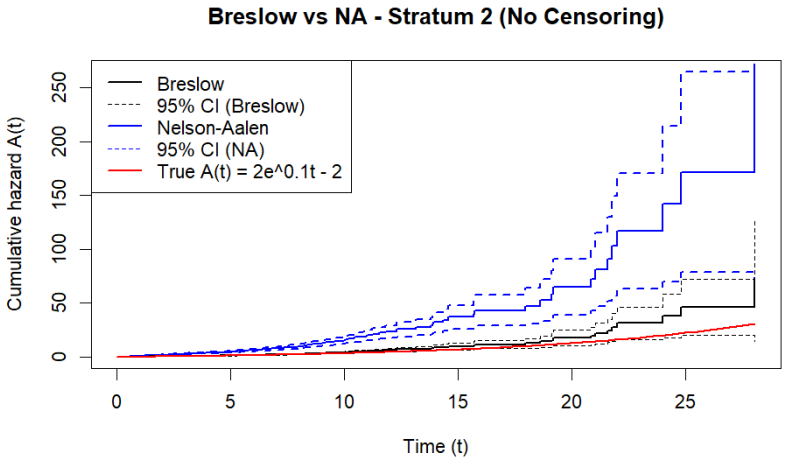

Figure 36: Nelson–Aalen and Breslow cumulative hazard estimates with 95% confidence intervals under a baseline cumulative hazard $A_{20}(t) = 2e^{0.1t} - 2$ for stratum 2, alongside the true cumulative hazard (red). Two covariates are used, of which one has a standard normal distribution, whereas the other has a discrete, symmetrical binomial distribution.

8.3 Breslow–Cox vs. Breslow–GBM

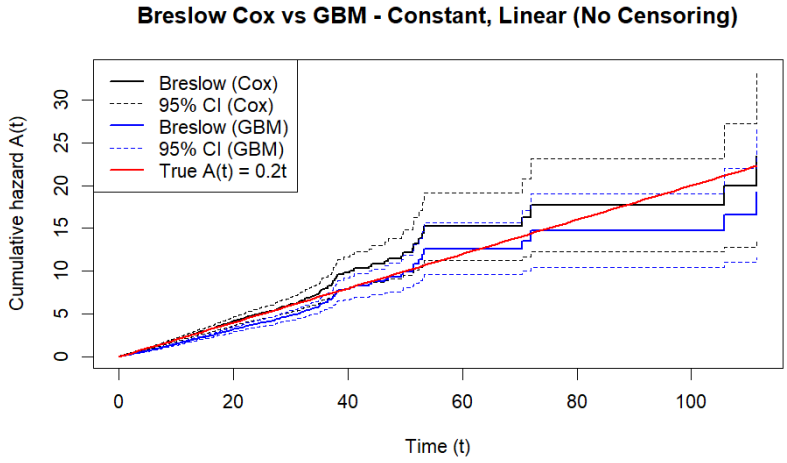

Figure 37: The Breslow cumulative hazard estimates with and without the aid of a GBM estimating $f(\mathbf{x}_i) = 0.5x_{i1} - 0.7x_{i2}$. Their respective 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$, alongside the true cumulative hazard (red).

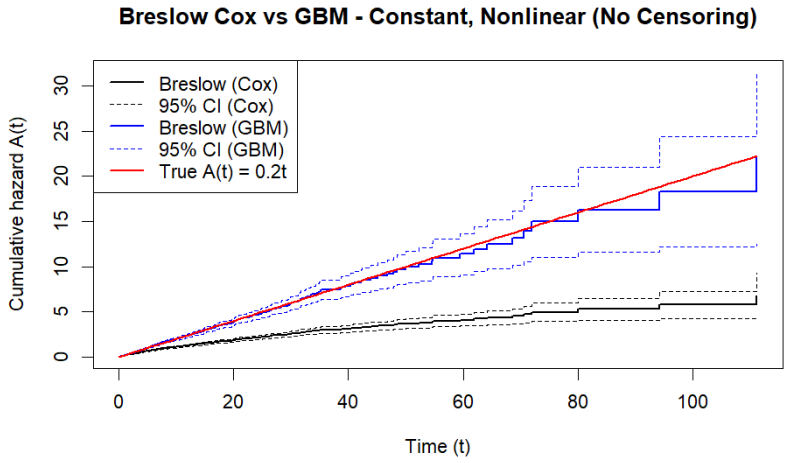

Figure 38: The Breslow cumulative hazard estimates with and without the aid of a GBM estimating $f(\mathbf{x}_i) = \sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$. Their respective 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 0.2t$, alongside the true cumulative hazard (red).

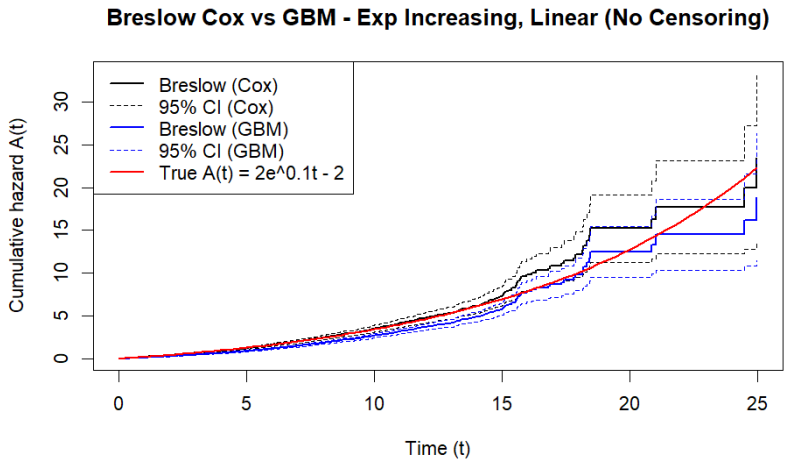

Figure 39: The Breslow cumulative hazard estimates with and without the aid of a GBM estimating $f(\mathbf{x}_i) = 0.5x_{i1} - 0.7x_{i2}$. Their respective 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$, alongside the true cumulative hazard (red).

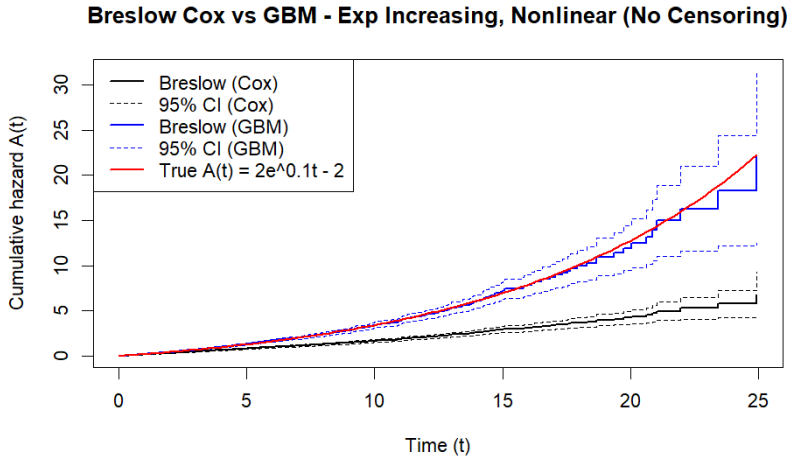

Figure 40: The Breslow cumulative hazard estimates with and without the aid of a GBM estimating $f(\mathbf{x}_i) = \sin(\pi x_{i1}) + 0.5(2x_{i2} - 1)$. Their respective 95% confidence intervals under a baseline cumulative hazard $A_0(t) = 2e^{0.1t} - 2$, alongside the true cumulative hazard (red).