# Assessment of Time-dependent Discriminative Ability

Yi He (s1684698)

First supervisor: Dr. Ir. Nan van Geloven (LUMC)

Second supervisor: Prof. Dr. Hein Putter (LUMC)

Third supervisor: Dr. Marta Fiocco (LUMC & MI)

MASTER THESIS

Defended on November $23^{rd}$, 2017

Universiteit Leiden
The Netherlands

Statistical Science

Leiden University Medical Center

**STATISTICAL SCIENCE**
**FOR THE LIFE AND BEHAVIOURAL SCIENCES**

**Abstract**

The area under the receiver operating characteristic (ROC) curve (AUC) is a commonly used measurement for the discriminative ability of a model. For the time to event variable in survival analysis the case and control sets will vary over time, thus a dynamic definition of AUC is required. We choose the dynamic AUC defined by incident true positive rate and dynamic false positive rate (I/D AUC) proposed by Heagerty and Zheng [6]. However, the difficulty to empirically obtain the incident true positive rate is hampering the estimation of dynamic AUC. Thus, several semi-parametric and non-parametric estimators are proposed. Heagerty and Zheng [6] proposed the semi-parametric estimation method based on Cox model. The non-parametric estimates using intermediate concordance measure with LOWESS smoothing is raised by van Houwelingen and Putter [14]. Based on the same intermediate concordance measure, Saha-Chaudhuri and Heagerty suggested to use locally weighted mean rank smoothing [10]. Recently, Shen et al proposed a semi-parametric method by adopting fractional polynomial to fit the dynamic AUC [12].

In this thesis, we compare the performance of these methods with different configuration in a series of simulations. The plain Cox methods is not recommended when the proportional hazards assumption is not satisfied. The Cox model with time-varying coefficients are relatively stable when the marker has a mediocre effect. For the non-parametric methods, a too wide span/bandwidth may lead to large bias, and a too narrow span/bandwidth may lead to unstable estimates, thus, the trade-off between the bias and the standard deviation has to be made. For fractional polynomial, adding extra fractional polynomial terms does not benefit the performance.

In addition, many researchers observed a decreasing trend of I/D AUC over time in their empirical studies [10][12][6], yet Pepe et al. held the opinion that the I/D AUC may be an increasing function over time [7]. We investigate the trend of I/D AUC under a Cox model and binary marker setting. However, we observe that under certain Cox models, the I/D AUC curve first increases then decreases, thus I/D AUC is not necessarily a decreasing function of time.

**Acknowledgements**

I wish to express my deepest gratefulness to my thesis supervisors Dr. Ir. Nan van Geloven and Prof. Dr. Hein Putter. The insightful weekly discussions during these six months helped me a lot. Without their guidance, advise and encouragement before the presentation, this thesis would not have been possible. I would also like to thank Dr. Marta Fiocco for helping me with the graduation procedure.

Finally, I wish to thank my internship supervisor Dr. Elise Dusseldorp for her support and encouragement throughout my final year. I want to thank my family and friends for accompanying me during my study.

# Contents

# Chapter 1

# Introduction

Nowadays, biomarkers become more accessible for diagnostic or prognostic uses. With a relatively large biomarker set, researchers would like to know which biomarker has the highest discriminative ability to predict whether a patient experiences an event in the future. To assess the discriminative ability of a biomarker or a set of biomarkers, a common choice is the area under the receiver operating characteristic curve (AUC). This idea is commonly used in logistic regression with a binary outcome and static predictors. However, the outcome variable used in survival analysis is usually a time-to-event variable. In addition, there may be longitudinal markers measured at different time points which are also time-dependent. Thus, the general ROC and AUC need to be extended to incorporate these time-dependent aspects of the data.

Pivotal work in this area has been done by Heagerty and Zheng[6]. They proposed several time-dependent discriminative measurements defined by combinations of incident/cumulative specificity and dynamic/static specificity. In this study, we will focus on the commonly used [7] time-dependent ROC and AUC formed by incident sensitivity and dynamic specificity. By definition, the evaluation of discriminative ability is based on the risk set at time $t$. The subjects who experience an event at time point $t$ serve as incident cases, and the event-free subjects until time $t$ serve as dynamic controls. Thus, this measure is a natural companion to hazard models [6]. Moreover, this measure offers a direct way to incorporate longitudinal markers (i.e. plug in the marker value at time $t$), while the measures defined by cumulative sensitivity need joint modeling with an extra linear mixed model [8].

Several estimation methods of incident/dynamic AUC (I/D AUC) were proposed in previous studies. Heagerty and Zheng (2005) proposed three semi-parametric estimation methods of I/D AUC based on Cox models [6]. Later in 2011, van Houwelingen and Putter suggested that, to obtain an estimate of time-dependent AUC, one can first estimate an intermediate concordance measure and then apply locally weighted smoothing (LOWESS) [14]. Saha-Chaudhuri and Heagerty came up with the locally weighted mean rank (WMR) smoothing instead of LOWESS smoothing based on the same intermediate concordance measure[10], while Shen et al. proposed a direct method by fitting the time-dependent AUC as a fractional polynomial of time [12]. As for evaluation of the methods, Schmid and Potapov (2012) compared different approaches (i.e. Heagerty and Zheng[6], Song and Zhou[13], Harrells C, etc.) of estimating concordance index based on I/D AUC by simulation studies [11]. However, the estimation of I/D AUC at different time points was not investigated in that study. In this thesis, we aim to inspect the performance

of the estimators at different time points.

Another interesting aspect mentioned in previous studies was the trend of I/D AUC over time. Heagerty and Zheng showed that a baseline model score had a higher discriminative ability at early time points than later time points in Veteran's Administration lung cancer data. The explanation offered in the paper was, for a time-varying health status measurement (i.e. Karnofsky score), the baseline marker values were not able to identify the future health status of a patient, which could lead to a substantial decline of the model's discriminative ability over time [6]. Similar results were found by Shen et al. when applying their method to evaluate the discriminative ability of AIDS biomarkers CD4 and CD8. The AUC curves of both markers showed a decreasing trend over time[12]. Yet another more theoretical comment made by Pepe et al. stated that the incident true positive rate might be a decreasing function of time for cancer screening, while anticipating the trend of dynamic false positive rate was more difficult since since it was related to both of the change of control group and the change of event detection properties[7].

In this paper, we will focus on the four I/D AUC estimation methods, namely Cox-model based estimation in [6], non-parametric estimations with LOWESS smoothing in [14], non-parametric estimations with locally WMR smoothing in [10], and fractional polynomial based estimation in [12]. The performance of these methods will be assessed under different simulation scenarios by the deviation of the simulation mean AUC($t$) from theoretical AUC($t$) and robustness measured by simulation standard deviations of the estimates. Thus, the first research question is, which method outperforms others under certain simulation scenarios in terms of bias and robustness, and is there any overall winner? Furthermore, we will explore the trend of I/D AUC by evaluating the derivatives of the theoretical I/D AUC expression. The second research question is, what trend can be expected of I/D AUC curve over time, does a decreasing trend of I/D AUC hold under common scenarios?

The paper is organized as follows. Chapter 2 contains the general definition of ROC and I/D ROC, a brief overview of the four methods and simulation schemes. The results of simulation studies are illustrated in Chapter 3. The investigation of the trend of I/D AUC is covered in Chapter 4. The fifth chapter involves general remarks of previous results, a discussion and some ideas for future work.

# Chapter 2

# Methodology

## 2.1 Notation

Assume we have $n$ subjects in the study. Let $T_i$ and $C_i$ denote the event time and the independent censoring time for subject $i$. The follow up time for subject $i$ is defined as $Z_i = \min(T_i, C_i)$, and the corresponding event indicator is defined as $\delta_i = \mathbf{1}(T_i \leq C_i)$. The at-risk indicator at time $t$ is defined as $r_i(t) = \mathbf{1}(Z_i \geq t)$. The subjects at risk at time $t$ form the risk set $R(t)$. We denote the size of the control set (i.e. the censored observations) at time $t$, $R_0(t)$, as $n_0(t)$ and the size of the case set (i.e. events) at time $t$, $R_1(t)$, as $n_1(t)$.

The marker of subject $i$ is denoted by $M_i$. Note that if we are interested in a model score rather than a single marker, the meaning of $M_i$ can be extended to $M_i = X_i^{'}\beta$, where $\beta$ is the coefficient vector from some parametric models (i.e. Cox model) and $X_i$ the marker values of subject $i$.

## 2.2 Definition of I/D ROC and AUC

The discriminative ability of a classifier can be accessed by the receiver operating characteristic (ROC) curve and the area under the curve (AUC). The ROC curve plots the true positive rate (TP, sensitivity) against the false positive rate (FP, 1 - specificity), under different cut-off values $c$ of the model score or marker $m$. Here is an example of classifying a binary outcome $A$ based on $m$. This classifier predicts $\hat{A} = 1$ when $m$ is larger than the threshold $c$ and $\hat{A} = 0$ when $m$ is smaller than the threshold $c$. Thus, the true positive rate, which is the probability of $\hat{A} = 1$ given $A = 1$, can be written as

$$\text{TP}(c) = \text{sensitivity}(c) = P(m > c \mid A = 1).$$

The false positive rate, which is the probability of $\hat{A} = 1$ given $A = 0$, can be written as

$$\text{FP}(c) = 1 - \text{specificity}(c) = 1 - P(m \leq c \mid A = 0) = P(m > c \mid A = 0).$$

For each cut-off $c$, we have a pair of coordinates [TP(c),FP(c)], which is a point in the TP - FP plane. The ROC curve is obtained by connecting the corresponding points for all possible cut-off values $c$. The ROC curve is then defined as $\{\mathrm{ROC}(p); 0 \leq p \leq 1\}$, with

$$\mathrm{ROC}(p) = \mathrm{TP}\left\{[\mathrm{FP}]^{-1}(p)\right\},$$

where $[\mathrm{FP}]^{-1}(p)$ denotes the cut-off $c_p$ that $\mathrm{FP}(c_p) = p$.

Thus, the area under the ROC curve (AUC) can be calculated by

$$\mathrm{AUC} = \int_0^1 \mathrm{TP}\left\{[\mathrm{FP}]^{-1}(p)\right\} dp.$$

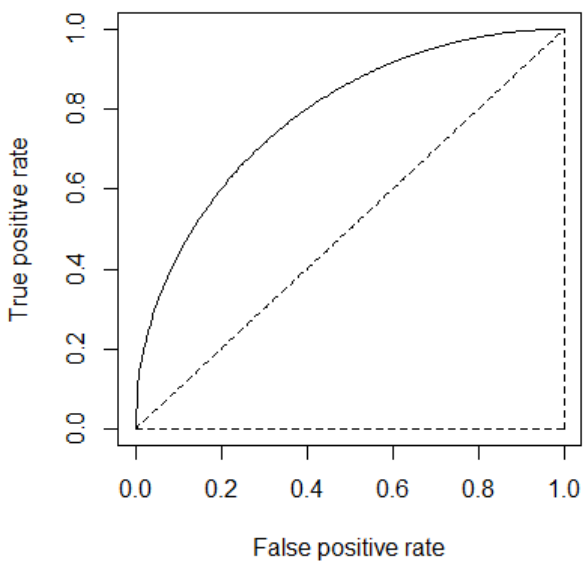


Figure 2.1: An example of the ROC plot. The area under the ROC curve is the AUC.

The value of AUC ranges from 0 to 1 (Figure 2.1). An AUC of 0.5 indicates the marker has no discriminative ability and an AUC larger than 0.5 indicates the marker outperforms random guessing. When the AUC equals 1, the marker perfectly separates the outcome. For AUC smaller than 0.5, the marker performs worse than random guessing. In this case, the predicted outcome can be reversed and the $\mathrm{AUC}_{new}$ becomes 1 - $\mathrm{AUC}_{old}$, thus the marker still has some discriminative ability.

In survival analysis, the outcome is a time to event variable, where the status of a subject can change over time. When estimating ROC and AUC under this circumstance, time-dependent versions of sensitivity and specificity are required. In this study, we are interested in the incident sensitivity /dynamic specificity (I/D) ROC and AUC defined by Heagerty and Zheng[6]. The incident true positive rate and the dynamic false positive rate are defined as follows

$$\mathrm{TP}_t^I = P(M_i > c \mid T_i = t) = \mathrm{sensitivity}^I(c, t);$$

$$\mathrm{FP}_t^D = P(M_i > c \mid T_i > t) = 1 - \mathrm{specificity}^D(c, t).$$

The incident cases are the subjects who experience an event at time $t$ ($T_i = t$), while the subjects

who had an event before time $t$ are left out. The dynamic controls are the event-free subjects until time $t$ (i.e. $T_i > t$). The subjects who serve as controls at time $t$ may become cases at later time points.

Thus, for $t \in T$, the I/D ROC is defined as

$$\mathrm{TP}_t^I \left\{ [\mathrm{FP}_t^D]^{-1}(p) \right\},$$

and the I/D AUC is defined as

$$\mathrm{AUC}(t) = \int_0^1 \mathrm{TP}_t^I \left\{ [\mathrm{FP}_t^D]^{-1}(p) \right\} dp = \int_{-\infty}^{\infty} \mathrm{TP}_t^I(c) \left| \frac{\partial \mathrm{FP}_t^D(c)}{\partial c} \right| dc.$$

## 2.3 Different methods of estimating I/D AUC

In this section, we will discuss four different methods of estimating I/D AUC, including two semi-parametric method and two non-parametric methods.

### 2.3.1 Semi-parametric - Cox model based estimation

Heagerty and Zheng proposed a semi-parametric estimation of time-dependent AUC based on Cox models [6]. Under the proportional hazards assumption, the Cox model can be written as $\lambda(t \mid M_i) = \lambda_0(t) \cdot \exp[M_i \cdot \gamma]$, where $\lambda(t \mid M_i)$ is the conditional hazard at time $t$, $\lambda_0(t)$ is the baseline hazard, $\gamma$ is the coefficient of the Cox model and $M_i$ is the model score or a single marker.

The weights $\pi_i(\gamma, t) = r_i(t) \cdot \exp(M_i \cdot \gamma) / \sum_j [r_j(t) \cdot \exp(M_j \cdot \gamma)]$ derived after fitting this model to data with the model score from any other models can be used to estimate the distribution of $M_i$ given the subject experiences an event at time $t$ [15]. Based on this, the incident sensitivity can be estimated by

$$\widehat{\mathrm{TP}}_t^I(c) = \hat{P}(M_i > c \mid T_i = t) = \sum_k \mathbf{1}(M_k > c) \pi_k(\gamma, t).$$

Estimation of dynamic specificity is easier, since there are more dynamic controls than incident cases for a specific time $t$ in survival data. The paper proposed an empirical estimator of dynamic specificity:

$$\widehat{\mathrm{FP}}_t^D(c) = \hat{P}(M_i > c \mid T_i > t) = \sum_k \mathbf{1}(M_k > c) \cdot r_k(t+) / \sum_k r_k(t+),$$

where $r_k(t)$ is the at risk indicator for control subject $k$ at time $t$, $r_k(t+) = \lim_{\delta \to 0} r_k(t + |\delta|)$, and $\sum_k r_k(t+)$ denotes the size of control set at time $t$.

However, the proportional hazards assumption is hardly ever satisfied in reality. A common way to deal with the violation of the proportional hazards assumption is adopting the time-varying coefficient model defined as $\lambda(t \mid M_i) = \lambda_0(t) \cdot \exp[M_i \cdot \gamma(t)]$. The dynamic specificity does not involve $\gamma(t)$, thus the definition stays the same.

The incident sensitivity is then defined as

$$\widehat{\mathrm{TP}}_t^I(c) = \sum_k \mathbf{1}(M_k > c)\pi_k(\gamma(t), t).$$

The time varying coefficient $\gamma(t)$ can be obtained by the smoothing of Schoenfeld residuals or by locally weighted maximum partial likelihood estimator (MPLE, local Cox). Note that the span of smoothing needs to be specified for both methods. More details about residual smoothing and locally weighted MPLE estimation can be found in [5] and [2].

For all possible values of c, we combine $\widehat{\mathrm{TP}}_t^I(c)$ with $\widehat{\mathrm{FP}}_t^D(c)$ to obtain a ROC curve. The AUC value is then calculated by numerical integration. The variance of estimates can be obtained by non-parametric bootstrapping. The algorithms proposed by the original paper [6] are available in the R package *risksetROC*.

### 2.3.2 Non-parametric - Locally weighted scatter plot smoothing

In [14], van Houwelingen and Putter proposed a non-parametric estimator of time-dependent AUC for untied data using locally weighted scatter plot smoothing (LOWESS). Rather than specifying sensitivity and specificity separately, this method first estimates an intermediate concordance measure defined as

$$\mathrm{A}(t) = \frac{\sum_{j \in R(t)} \mathbf{1}(M_j < M_i) + 0.5 \sum_{j \in R(t)} \mathbf{1}(M_j = M_i)}{n_0(t)},$$

where the subject $i$ dies at time $t$. When event times contain ties, the intermediate concordance measure can be extended to:

$$\mathrm{A}(t) = \frac{\sum_{j \in R_0(t)} \sum_{i \in R_1(t)} \mathbf{1}(M_j < M_i) + 0.5 \sum_{j \in R_0(t)} \sum_{i \in R_1(t)} \mathbf{1}(M_j = M_i)}{n_0(t) \cdot n_1(t)}.$$

This intermediate measure is the proportion of controls who have smaller or equal marker values than the incident case $i$ at time $t$ among the risk set at time $t$. Once the intermediate results $\mathrm{A}(t)$ are obtained, LOWESS is adopted to estimate time-dependent AUC. Instead of performing regression based on the entire dataset, LOWESS fits a local linear polynomial to each subset of data generated by a moving window. The moving window (span) is defined as a proportion of neighboring time points around time $t$. Thus, this method provides a good visualization of local patterns in the data. However, the explicit expression of regression function is not available under this non-parametric setting. Details about LOWESS can be found in [3]. The variance of the estimates can be obtained by bootstrapping. This algorithm (untied version) is implemented in R package *dynpred*, using the default span of LOWESS.

### 2.3.3 Non-parametric - Locally weighted mean rank smoothing

Another non-parametric estimator of time-dependent AUC using locally weighted mean rank smoothing was proposed by Saha-Chaudhuri and Heagerty in [10]. This method first estimates the same intermediate concordance measure $\mathrm{A}(t)$ for all event time $t$ as van Houwelingen and

Putter. Then the weighted mean rank (WMR) of a specified neighborhood around $t$ is calculated. Denote the neighbor around $t$ as $N_t(h_n) = \{t : |t - t_j| < h_n\}$, where the span $h_n$ controls the size of the smoothing window. The WMR is simply defined as the mean of $\mathrm{A}(t)$ in neighborhood $[t - h_n, t + h_n]$:

$$\mathrm{WMR}(t) = \frac{1}{|N_t(h_n)|} \sum_{t_j \in N_t(h_n)} \mathrm{A}(t_j),$$

where $|N_t(h_n)|$ denotes the size of the neighborhood. The $\mathrm{WMR}(t)$ can also be generalized to $\widehat{\mathrm{WMR}}(t) = \sum_j K_{h_n}(t - t_j) \cdot \mathrm{A}(t_j)$, where $K_{h_n}(t - t_j)$ denotes an arbitrary standardized kernel function that sums to 1.

The paper illustrated the asymptotic distribution and variance estimation of $\mathrm{WMR}(t)$. The asymptotic optimal span $h_n^{opt}$ can be determined by optimizing the mean square error of $\mathrm{WMR}_{h_n}(t)$. However, the implementation of this algorithm is not available in R or other statistical software.

### 2.3.4 Semi-parametric - Fractional polynomial

Shen et al. proposed a semi-parametric method using fractional polynomials to model $\mathrm{AUC}(t)$ as a function of time [12]. Compared to conventional polynomials, fractional polynomials have a larger variety of shapes that can mimic most functions in reality [9]. Since AUC values are bounded by 0 and 1, we need to extend this range to $[-\infty, \infty]$ by a link function $\eta(\cdot)$ (i.e. logistic link). The model of transformed $\mathrm{AUC}(t)$ as function of $t$ can be written as

$$\eta(\mathrm{AUC}(t)) = \sum_{k=0}^{K} \beta_k \cdot t^{(p_k)},$$

where $\beta_k$ is the regression coefficient of the $k_{th}$ polynomial and $t^{(p_k)}$ is defined for real-valued power $p_k$ as

$$t^{(p_k)} = \begin{cases} t^{p_k} & \text{if} \quad \mathrm{p_k} \neq 0 \\ ln(t) & \text{if} \quad \mathrm{p_k} = 0 \end{cases}.$$

At an event time $t$, we count the concordant and discordant cases as

$$c_1(t) = \sum_j \mathbf{1}\{j : M_i > M_j, T_i = t, j \in R_0(t), i \in R_1(t)\};$$

$$c_2(t) = \sum_j \mathbf{1}\{j : M_i \leq M_j, T_i = t, j \in R_0(t), i \in R_1(t)\}.$$

Note that the counts $c_1(t)$ follows a binomial distribution with success probability $\mathrm{AUC}(t, \beta)$. Thus, the model can be estimated by maximizing a pseudo partial-likelihood defined as

$$L(\beta) \propto \prod_t \mathrm{AUC}(t, \beta)^{c_1(t)} [1 - \mathrm{AUC}(t, \beta)]^{c_2(t)}$$

The original paper chose $p_1, ..., p_k = \{-2, -1, -0.5, 0, 0.5, 1, 2\}$ as suggested in [9]. The asymptotic variance is given in the paper and the implementation of fitting polynomials is available as R code in supplementary material [12]. However, the implementation of variance calculation was not provided.

### 2.3.5 Configuration of methods

In this paper, we will evaluate the performance of the methods mentioned above with different specifications. The details can be found in Table 2.1.

Table 2.1: Methods overview.

| Cox model based estimation | LOWESS | WMR | Fractional polynomial |
|---|---|---|---|
| Plain Cox | span=0.01 | span=0.05 | degree = {-2,-1,-0.5,0,0.5,1,2}** |
| Local Cox (span=0.35*) | span=0.33 | span=0.35* | degree = {-3,-2,-1,-0.5,0,0.5,1,2,3} |
| Residual smoothing (span=0.35*) | span=0.66** | span=1.5 | |

*: the suggested span $n^{-0.2}$, where $n = 200$ in the simulation.
**: the default or suggested setting.

## 2.4 Simulation scenarios

### 2.4.1 Bivariate normal distribution based scenarios

The performance of the four different methods will be evaluated under a series of simulation scenarios. A commonly used distribution for data generation is bivariate normal distribution with mean $\begin{bmatrix} \mu_1 \\ \mu 2 \end{bmatrix}$ and covariance matrix $\begin{bmatrix} \rho\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \rho\sigma_2^2 \end{bmatrix}$ denoted by $\mathrm{N}_2[\mu_1, \mu_2, \sigma_1, \sigma_2, \rho]$ [12][10][6]. Here, log event times $\log(T)$ and marker $M$ jointly follow $\mathrm{N}_2[\mu_1, \mu_2, \sigma_1, \sigma_2, \rho]$. Thus, the event times follow a Log-normal distribution with mean $\mu_2$ and standard deviation $\sigma_2$, and the marker follows a normal distribution $N(\mu_1, \sigma_1)$. In this study, we generate the data from $\mathrm{N}_2[0, 0, 1, 1, \rho]$. The hazard of log-normal event time is shown in Figure 2.2, most of the events will happen in the beginning.

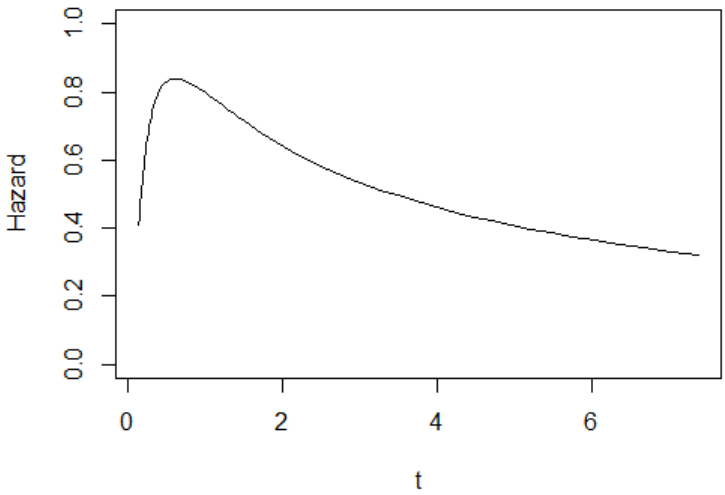


Figure 2.2: Hazard of standard log-normal distribution

The correlation between event time and marker can be easily controlled by $\rho$. In this paper, we consider a relatively high correlation $\rho = -0.7$ and a relatively low correlation $\rho = -0.3$.

Here, the negative correlation coefficients indicate that the subjects with higher marker values would experience an event earlier than the subjects who have lower marker values. Figure 2.3 shows the relationship between marker and event time generated from $\mathrm{N}_2[0, 0, 1, 1, \rho]$. According to the figure, the proportional hazards assumption does not hold under these settings.

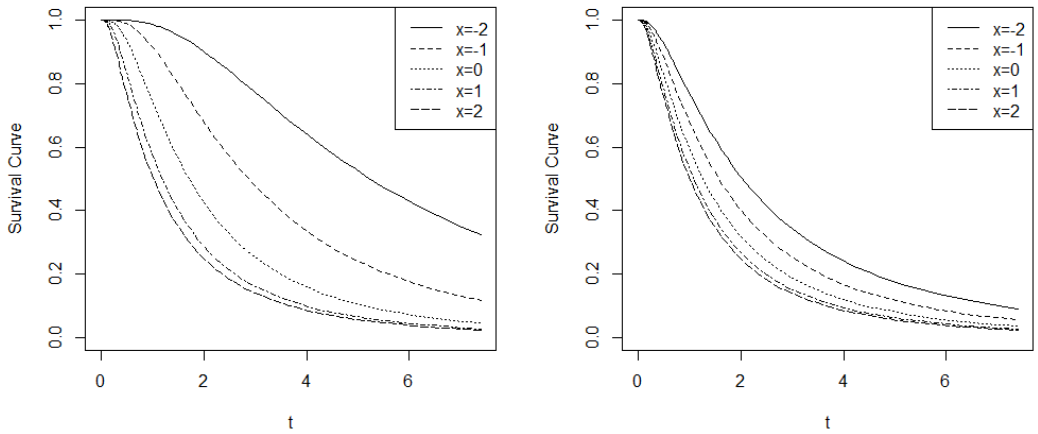


Figure 2.3: Relationship between marker and event time generated from $\mathrm{N}_2[0, 0, 1, 1, \rho]$. Left panel: $\rho = -0.7$, right panel: $\rho = -0.3$.

Once the joint distribution of log event time and marker is obtained, we can calculate the theoretical value of I/D AUC. By definition[6], the theoretical incident sensitivity and dynamic specificity under $\mathrm{N}_2[0, 0, 1, 1, \rho]$ can be calculated by

$$\text{sensitivity}^I(c, t) = P(M_i > c \mid T_i = t) = \Phi[\frac{\rho \log(t) - c}{\sqrt{1 - \rho^2}}];$$

$$\text{specificity}^D(c, t) = P(M_i \leq c \mid T_i > t) = 1 - \frac{S_2^N[c, \log(t), \rho]}{\Phi[-\log(t)]},$$

where $S_2^N$ denotes the joint survival distribution of $\mathrm{N}_2[0, 0, 1, 1, \rho]$ and $\Phi[x]$ denotes the cumulative density of standard normal distribution $N[0, 1]$. Then the theoretical AUC values can be obtained by numerical integration. Figure 2.4 shows the theoretical AUC(t) curves under different $\rho$'s.

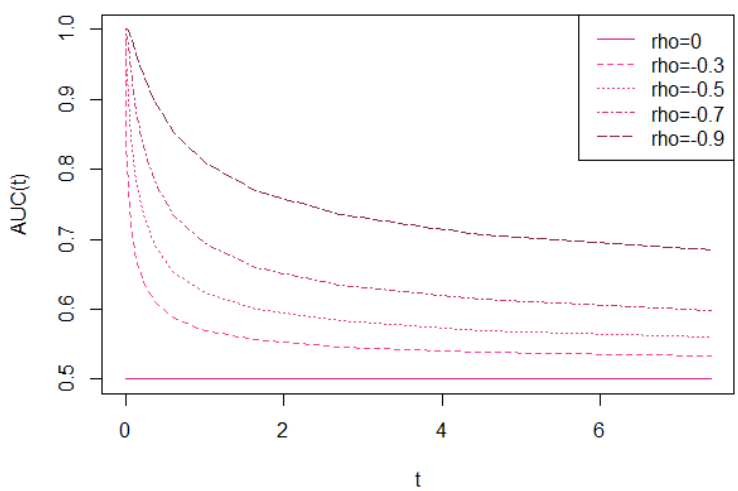


Figure 2.4: Theoretical AUC values under bivariate normal distribution with different $\rho$'s.

Furthermore, different censoring distribution and percentage of censoring can influence the estimation of AUC. Different censoring distributions have an impact on the density of event times at different time ranges. The percentage of censoring controls the number of events in the data. When the data are sparse, the estimations of AUC may contain larger bias and variance since less information and more interpolations are involved. In this study, exponential, log-normal, Weibull and uniform distributions under different percentages of censoring $\{0\%, 20\%, 40\%, 60\%, 80\%\}$ are used for generating censoring times.

For each scenario, 1000 datasets ($n$=200) are generated. The AUC($t$) values are evaluated under a series of log time points {-2,-1.5,-1,-0.5,0,0.5,1,1.5,2}, which correspond to {-2,-1.5,-1,-0.5,0,0.5,1,1.5,2} standard deviation(s) from mean 0. The mean over 1000 estimates will be compared to the theoretical AUC values in terms of bias and the robustness of the methods will be examined by the standard deviations of 1000 estimates.

### 2.4.2 Cox model based scenarios

Apart from the log-normal hazard, we will include other hazard shapes, i.e. constant, increasing and decreasing hazards in this study. Moreover, the Cox model based method requires proportional hazards for estimating the true positive rate, which is not satisfied under the bivariate normal distribution. The data are generated from the Cox model $\lambda(t \mid X = x) = \lambda_0(t)\exp(\beta X)$ with Weibull baseline hazards. The cumulative hazard function of the Weibull distribution is $\Lambda_0(t) = \alpha t^{\gamma}$ and the hazard function is $\lambda_0(t) = \alpha\gamma t^{\gamma-1}$. Different hazard shapes can be obtained by varying parameter $\gamma$. In this study, we choose $\gamma = 0$ for an exponential or constant hazard, $\gamma = 0.75$ for a decreasing hazard and $\gamma = 1.25$ for an increasing hazard, while $\alpha$ is set to 2.

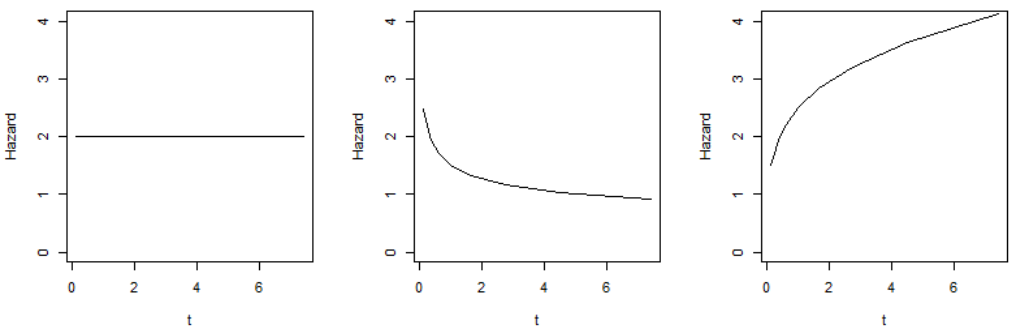


Figure 2.5: Different hazard shape of Cox model based data. Left panel: $\gamma = 1$, middle panel:$\gamma = 0.75$, right panel:$\gamma = 1.25$.

To generate the data from Cox model, we first sample the marker $x$ from a standard normal distribution then we sample the survival probability $u$ from a uniform distribution U(0,1). Then, based on $S(t \mid X = x) = \exp[-\Lambda_0(t) \cdot \exp(\beta x)]$, the survival time $t$ can be calculated by $t = [\frac{-\log(u)}{\alpha \cdot \exp(x \cdot \beta)}]^{1/\gamma}$. The marker values are generated from a standard normal distribution as the previous scenarios. The $\beta$'s are 0.4 and 1.25 approximately in accordance with $\rho = -0.3$ and $\rho = -0.7$ under bivariate normal settings. Figure 2.6 illustrates the relationship between event time and different marker values. Note that the event time range also varies in relation to hazard shape.

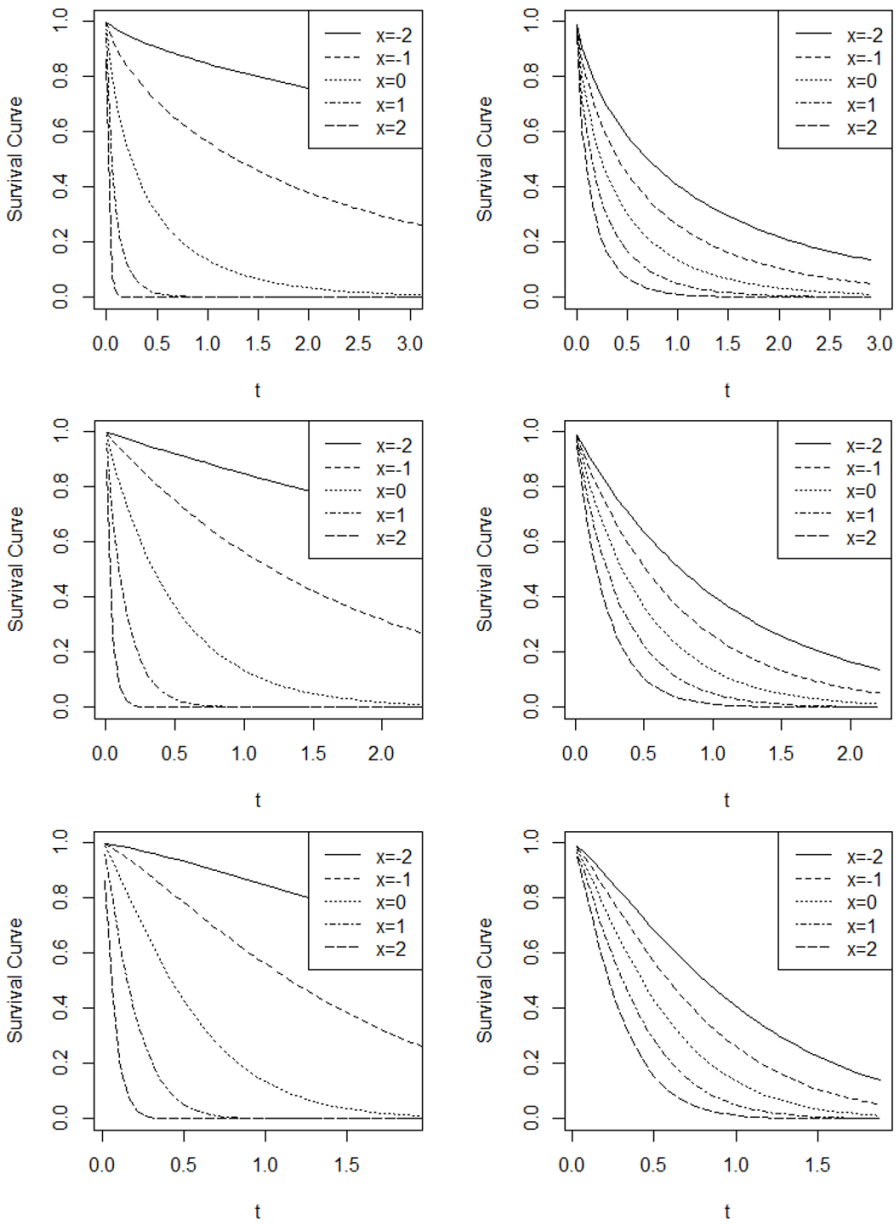


Figure 2.6: Relationship between marker and event time generated from Cox model. For a fair visualization, time ranges from the $2.5^{th}$ to the $97.5^{th}$ percentiles of $\beta = 0.4$ data. Left panels: $\beta = 1.25$, right panels: $\beta = 0.4$. Top panels: decreasing hazard with $\gamma = 0.75$. Middle panel: exponential (constant) hazard with $\gamma = 1$. Bottom panel: increasing hazard with $\gamma = 1.25$

The theoretical I/D AUC at different time points can be obtained by numerical integration once we have the true positive and false positive rates. Under these scenarios, the incident true positive rate is $P(M_i > c \mid T_i = t) = \frac{\int_c^\infty f(t|x)g(x)dx}{f_T(t)}$, and the dynamic false positive rate is $P(M_i > c \mid T_i > t) = \frac{\int_c^\infty S(t|x)g(x)dx}{S_T(t)}$ [1], where $f_T(t)$ and $S_T(t)$ are density and survival functions at $t$, and $g(x)$ is the density function of the marker.

Similar to previous scenarios, we create censoring patterns from different censoring distributions and percentages. For Cox model based data, we choose exponential distribution and uniform distribution with corresponding parameters that ensure there are 0%, 20% and 60% censoring subjects in the data. For each scenario, we generate 1000 datasets of size 200 ($n = 200$). We evaluate the AUC values under a series of time points determined by the empirical percentiles {2.3%, 6.7%, 15.9%, 30.9%, 50.0%, 69.1%, 84.1%, 93.3%, 97.7%} of the event times. In terms of bias, we will compare the mean over 1000 estimates with the true AUC values. The standard deviations of 1000 estimates are also calculated to assess the robustness of the methods.

# Chapter 3

# Results

In this chapter, we present the results from different simulation scenarios. In the first section, we illustrate the results of simulation studies and discuss the influence of censoring pattern and strength of correlation between log time and marker based on bivariate normal distribution based scenarios, and the influence of hazard shape based on Cox model based scenarios. In the second section, we provide a discussion about the different configurations (Table 2.1) of each method and their computation time. To visualize and evaluate the performance of the methods, we plot the mean of the estimates over 1000 simulated datasets and the corresponding standard deviation against time for each scenario.

## 3.1 Bivariate normal distribution based scenarios

We first illustrate the results of the bivariate normal based scenarios with $\rho = -0.7$. Figure 3.1 shows the mean and standard deviation of AUC($t$) estimates from the six methods under their default or suggested settings with 0% censoring. It can be observed that the true AUC($t$) values have a decreasing trend over time. With a large correlation $\rho = -0.7$, the true values are relatively high for early time points (i.e. when $t = 0.135$, the true AUC($t$) is 0.884). Compared to the true values, for parametric methods, the plain Cox estimates fail to pick up the trend of AUC($t$) over time, while other methods, namely local Cox, residual smoothing, and fractional polynomial, present a relatively satisfactory result yet with small deviance at later time points. For non-parametric methods, the non-parametric LOWESS overestimates the true AUC($t$) over time. The non-parametric WMR estimates, notwithstanding the underestimation at the beginning, keep aligned with the true values. In terms of simulation standard deviation, the Cox based methods have the smallest standard deviations, followed by the non-parametric LOWESS and fractional polynomial with larger standard deviations, while for non-parametric WMR, the standard deviations are non-negligible.

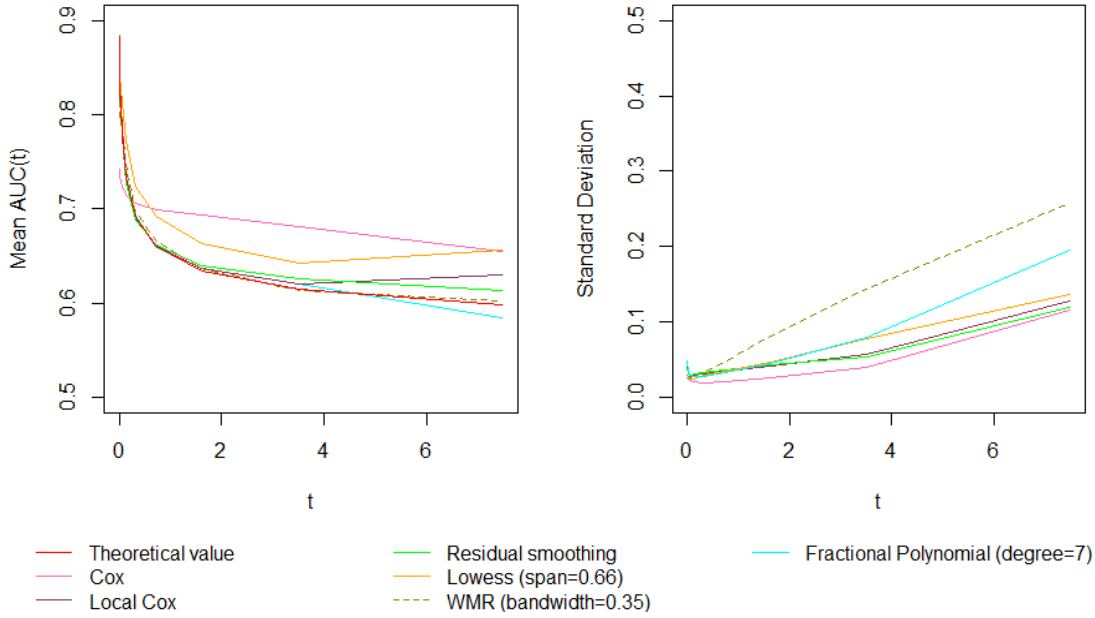


Figure 3.1: The performance of the six algorithms under simulation scenario 0% censoring, $\rho = -0.7$. Left panel: simulation mean of estimated AUC($t$). Right panel: simulation standard deviation of estimated AUC($t$).

### 3.1.1 Influence of censoring patterns

In this subsection, we illustrate the influence of censoring percentages and investigate the impact of different censoring distributions on the estimates. Generally, with an increasing censoring percentage, the number of events decreases. According to the defination, we can only obtain the I/D AUC estimates at the event time points, thus with less event times, the information provided by the data may be insufficient. We would like to know whether these methods are able to recover the trend of I/D AUC values with limited information.

We present the bivariate normal scenarios with exponential censoring to explain the effect of censoring percentage since the results under different censoring distributions do not differ much for 20% and 40% censoring. Comparing Figure 3.1 and 3.2, when the censoring percentage increases from 0% to 40%, we observe that the bias of all methods shows up earlier, and the magnitude of the bias is larger. The standard deviations of all methods also show a slightly increasing trend with an increasing censoring percentage. There are two possible explanations. Firstly, the AUC($t$) is calculated at each event time point, the sparsity of events would lead to less estimates of AUC($t$), thus large bias and larger standard deviations. Secondly, not all of the simulation datasets contain observed time until the later time points we evaluate, when averaging through a smaller set of estimates, the bias and the standard deviations may also go up. For Cox based methods, the estimates from plain Cox show the largest bias over time, while local Cox and residual smoothing retain a good fit at the beginning yet deviate heavily from the true values at later time points. Notwithstanding the bias, these methods are relatively stable since the standard deviations are small. The fractional polynomial shows evidently increasing standard deviations with the increasing censoring percentage. The non-parametric WMR estimates show a small bias but large standard deviations, while the non-parametric LOWESS overestimates the true value as before even though its standard deviations are relatively small.

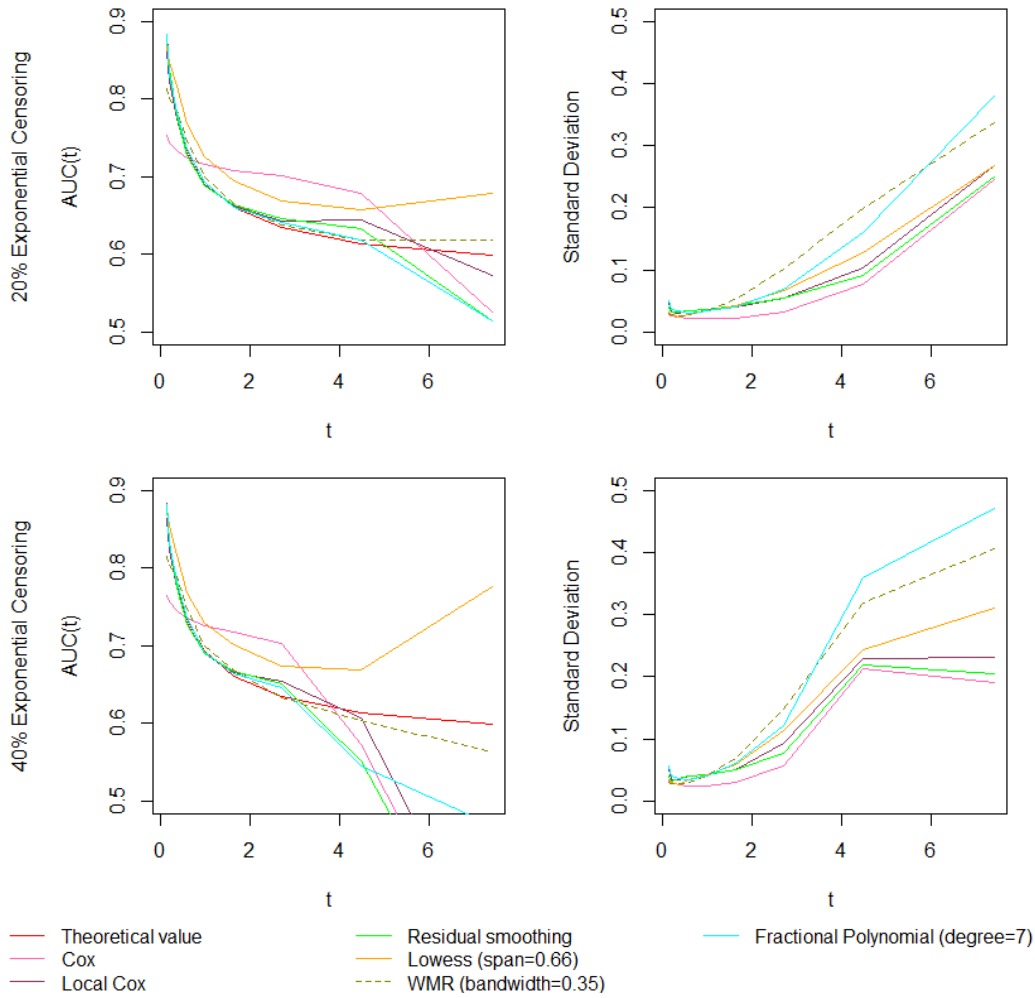


Figure 3.2: The results of the scenario under 20% and 40% exponential censoring with $\rho = -0.7$. Left panel: mean AUC($t$) over 1000 simulated datasets. Right panel: simulation standard deviation over 1000 simulated datasets.

For higher censoring percentages (i.e. 60%), we notice that the observed time ranges differ a lot. For each scenario, we show the survival curves obtained by one simulation dataset as an example (Figure 3.3). Under uniform censoring, only time points before $t = 1.6$ are observed, while for Weibull censoring, there are no events or censors happening after $t = 1$. As for log-normal and exponential censoring, the data contain few observed event times until $t = 3$.

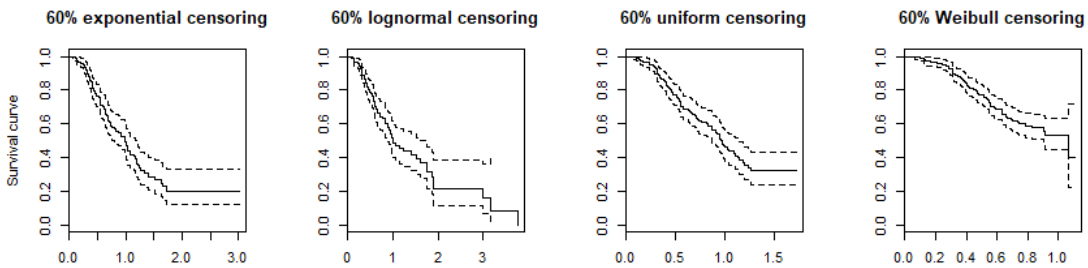


Figure 3.3: The survival curves obtained from one simulation dataset under 60% censoring with different censoring distributions.

From Figure 3.4, the performances of these methods under different censoring distributions are similar to previous scenarios (Figure 3.1 and 3.2). The estimates from plain Cox and non-parametric LOWESS have the largest bias, yet with smaller standard deviations over time. The local Cox, residual smoothing and fractional polynomial show an ideal fit for early time points, while the fractional polynomial has larger standard deviations. The non-parametric WMR shows a mediocre performance that the estimates contains moderate bias over time.

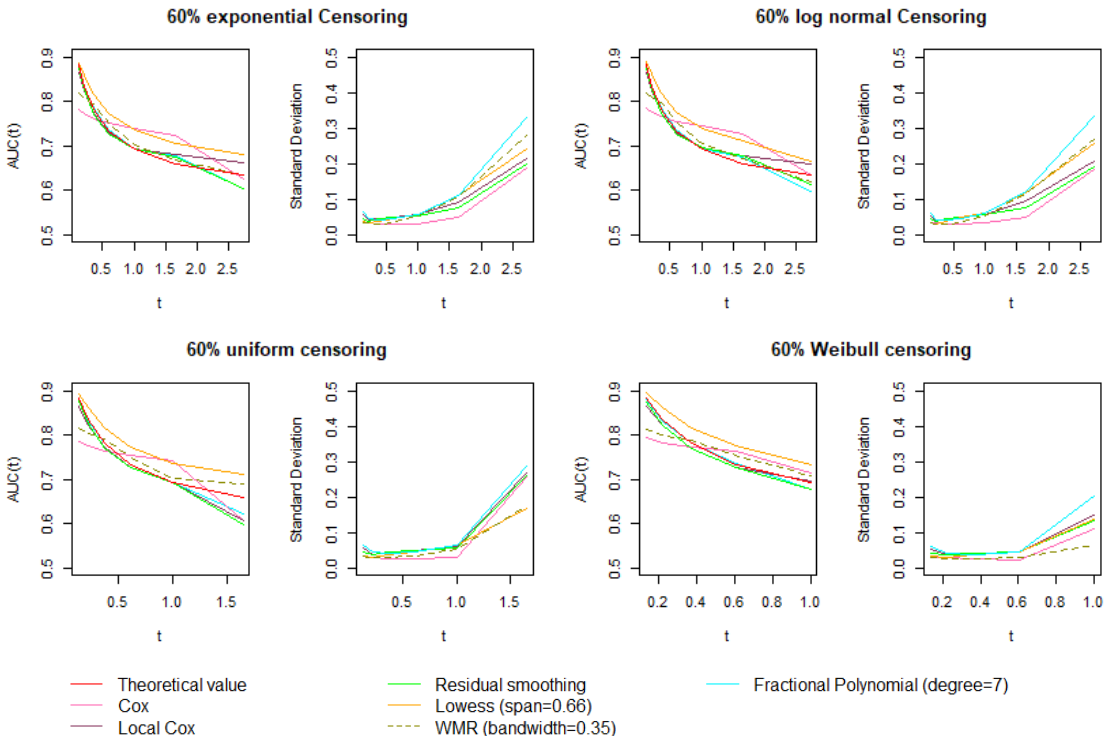


Figure 3.4: The results of simulation scenarios 60% censoring with $\rho = -0.7$. From left to right, upper to bottom: 60% exponential censoring, 60% log normal censoring, 60% Weibull censoring and 60% uniform censoring.

In conclusion, the censoring distributions and censoring percentages would influence the density of event times and the observed time range of the data. The data with censoring patterns generated from uniform and Weibull distributions tend to have shorter observed time ranges than log normal and exponential distributions. With an increasing censoring percentage, the bias of estimates shows up earlier with a larger magnitude and their simulation standard deviations also increase slightly. However, we observe similar performance of the methods under different censoring distributions. For earlier time points, the local Cox and residual smoothing outperform other methods with small bias and standard deviations, while for later time points, there is no sufficient evidence that an overall winner would exist. The plain Cox and non-parametric LOWESS (span=0.66) show highly biased results, although with low standard deviations, they are still not acceptable under bivariate normal scenarios with different censoring patterns.

### 3.1.2 Influence of strength of correlation

We consider the correlation $\rho$ between log event time $\log(T)$ and marker $M$ as an indicator of effect size. In this study, $\rho = -0.7$ and $\rho = -0.3$ are used to generate the bivariate normal simulation data. Figure 3.5 shows the results of the scenario without censoring under $\rho = -0.3$. Compared to $\rho = -0.7$ (Figure 3.1), the theoretical AUC values are much lower, which is in accord with our expectation. The Cox based methods show larger bias than $\rho = -0.7$, while non-parametric WMR and fractional polynomial have the smallest bias yet the largest standard deviations. The performance of non-parametric LOWESS is improved with a smaller overestimation. As shown in Figure 3.6, under 20% and 40% exponential censoring, all three Cox based methods heavily deviate from the true values, while other methods fail to balance their bias and standard deviations. Thus, when the correlation between log event time and marker is weak, Cox based methods are not recommended, while other methods need further evaluation under different configurations.

Other results under different censoring patterns of $\rho = -0.3$ can be found in Appendix A.2. Ordinarily, with different censoring patterns, the standard deviations of simulations under $\rho = -0.3$ are slightly higher at the beginning than that of $\rho = -0.7$, and the mean AUC($t$) shows larger deviances from the theoretical values.

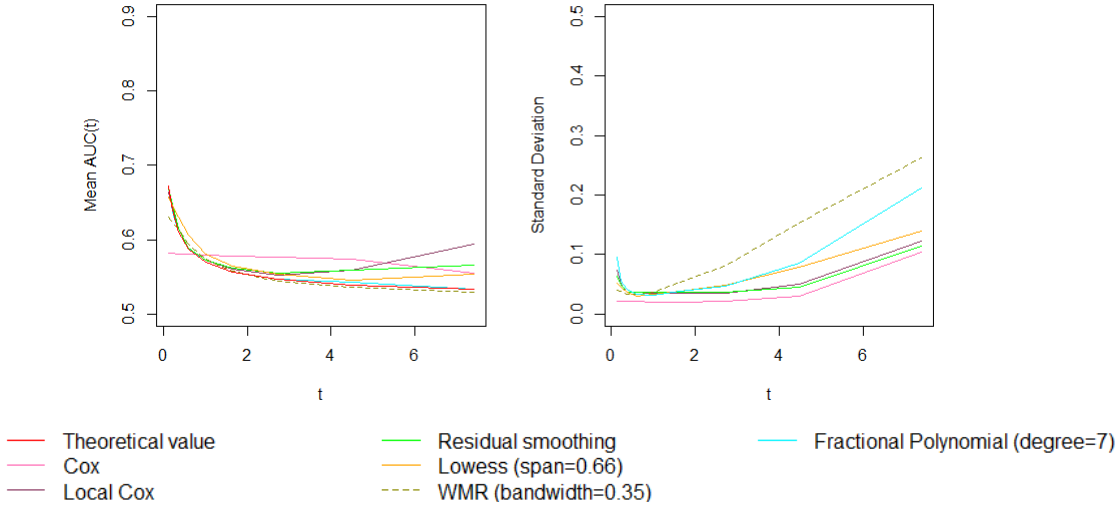


Figure 3.5: The performance of the six algorithms under simulation scenario 0% censoring with $\rho = -0.3$. Left panel: simulation mean of estimated AUC($t$). Right panel: simulation standard deviation of estimated AUC($t$).

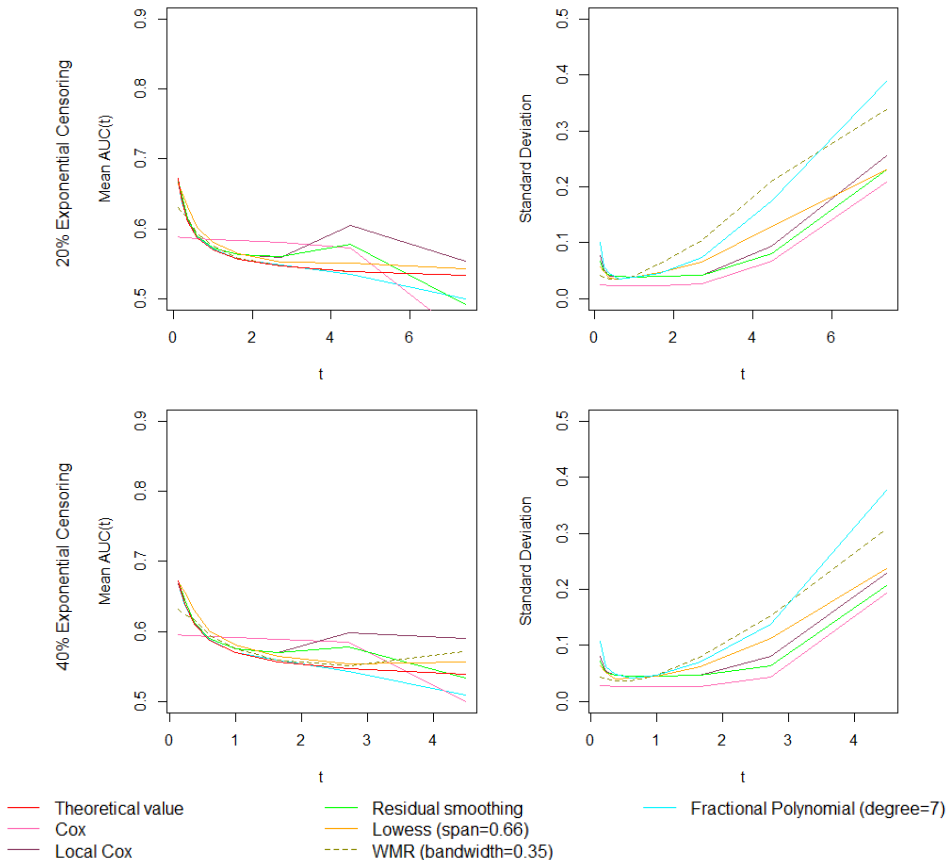


Figure 3.6: The results of the scenario under 20% and 40% exponential censoring with $\rho = -0.3$. Left panel: mean AUC($t$) over 1000 simulated datasets. Right panel: simulation standard deviation over 1000 simulated datasets.

## 3.2 Cox model based scenarios

In this study, apart from the bivariate normal distribution based scenarios, we also include Cox model based scenarios with Weibull baseline hazard ($\lambda_0(t) = \alpha\gamma t^{\gamma-1}$). By varying the parameter $\gamma$ of Weibull hazard function, we generate data from constant hazard, increasing hazard and decreasing hazard. Different hazard shapes would influence the observed time range. For the increasing Weibull hazard with $\beta = 1.25$, the time interval [0, 3.2] contains 95% of event times, while for constant hazard it is [0, 4.3], and for the decreasing hazard it is [0, 7.0].

The methods show similar performance under different simulation scenarios. Here we present the scenarios with $\beta = 1.25$ and 20% exponential censoring in Figure 3.7. The theoretical AUC curves under different scenarios all show a decreasing trend over time. With decreasing $\gamma$'s, for later time points, the bias and the standard deviations of the estimates increase. The intuition behind is, more events will happen at the beginning of the follow-up, which results in the sparsity of events for later time points. The plain Cox estimates perform better since the proportional hazards assumption is satisfied under these scenarios, and the results are similar to residual smoothing. However, the non-parametric LOWESS still overestimates the true values and the non-parametric WMR is unable to follow the trend of the true values for earlier time points. The estimates from fractional polynomial contains acceptable bias but large standard deviations.

Although the local Cox shows an ideal fit under 20% exponential censoring with increasing and constant hazards, this method gives numerical problems when there is local perfect separation in the data (Appendix B, Figure B.1, 0% censoring, decreasing hazard, $\beta = 1.25$). Thus, in terms of bias and standard deviation, residual smoothing and plain Cox could be recommended under the scenario where the proportional hazards assumption holds.

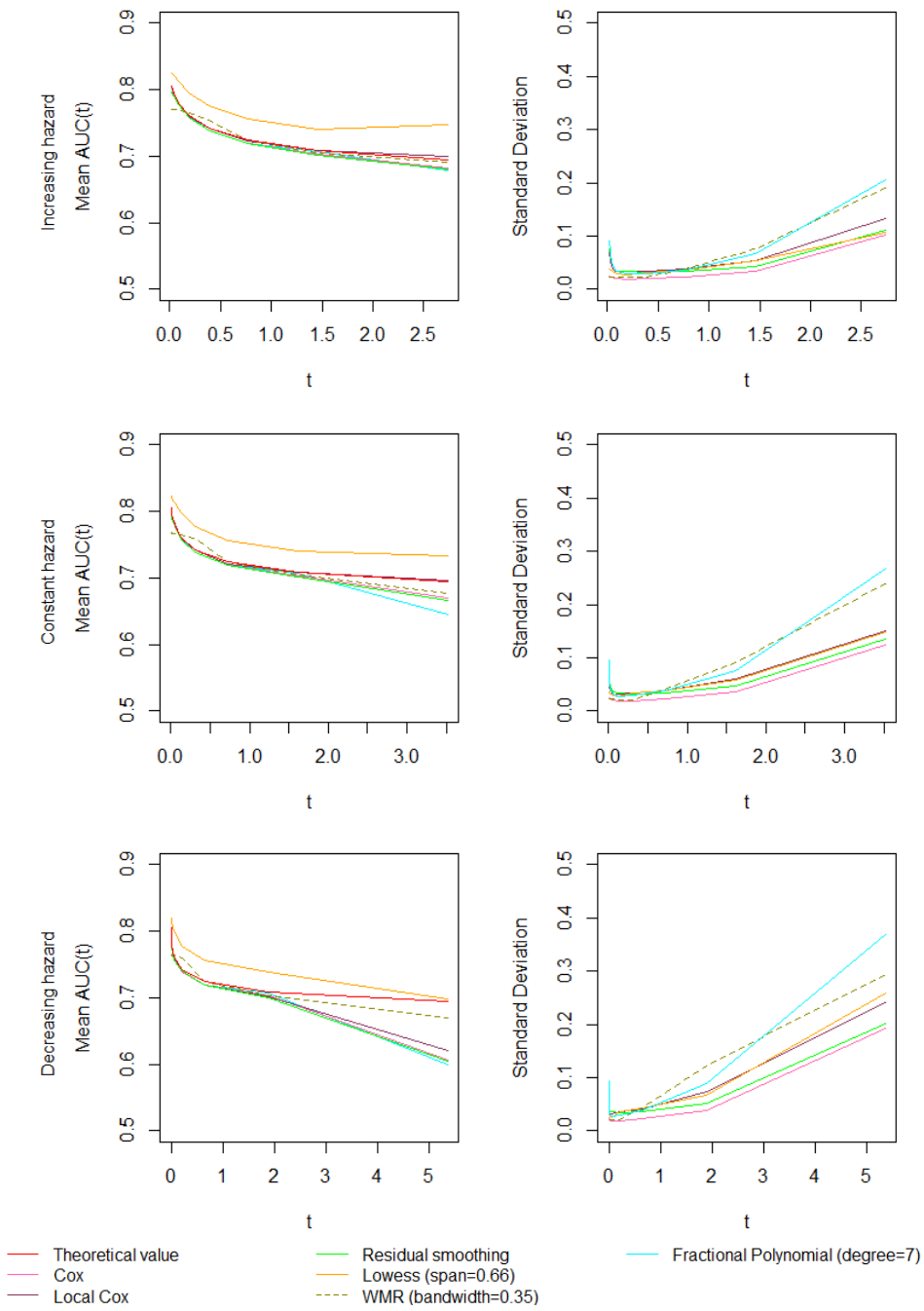


Figure 3.7: The results of simulation scenarios of different hazard shapes under 20% exponential censoring, $\beta = 1.25$. From top to bottom: increasing Weibull hazard ($\gamma = 1.25$), constant (exponential) hazard, decreasing Weibull hazard ($\gamma = 0.75$). Left panel: the mean AUC($t$) over 1000 simulated datasets. Right panel: the simulation standard deviations.

## 3.3 In-depth results of different methods

In this section, we illustrate the results from different configurations of each methods. Since the results under different censoring patterns and hazard shapes do not differ much, we mainly show the mean and standard deviations of the estimates from $\rho = -0.7$ with 20% uniform censoring under bivariate normal distribution and $\beta = 1.25$ with 20% uniform censoring under Cox model.

### 3.3.1 Semi-parametric Cox model based estimation

We adopt the suggested span $n^{-0.2}$ for local Cox and residual smoothing, where $n$ is the sample size per dataset. As mentioned in the original paper [6], the violation of the proportional hazards assumption would lead to biased estimates of the plain Cox algorithm. Accordingly, when the proportional hazards assumption is not satisfied (Figure 3.8, upper panel, green line), the estimates from plain Cox contain the largest bias, while under the Cox model based scenario, the estimates from plain Cox show a better behavior (Figure 3.8, bottom panel, green line). In contrast, the local Cox (orange line) and residual smoothing (blue line) relax the assumptions by adopting time-varying coefficients, and the results of these two algorithms are much better than the plain Cox. With a slight loss in standard deviation compared to the plain Cox, one could achieve a much smaller bias by using one of these two methods.

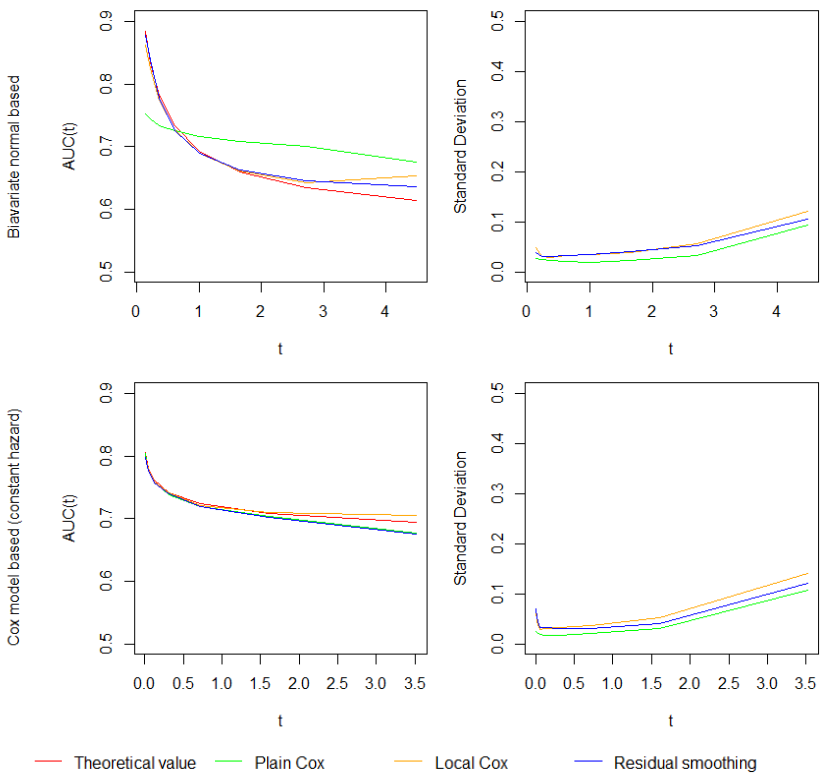


Figure 3.8: The performance of Cox model based algorithms under simulation scenario 20% uniform censoring with $\rho = -0.7$. From left to right: log normal hazard and constant hazard. Upper panel: simulation mean of estimated AUC($t$). Bottom panel: simulation standard deviation of estimated AUC($t$).

The plain Cox, as mentioned before, requires proportional hazards, thus under the cases where the assumption does not hold, this method might be inappropriate. The local Cox shows highly biased results under the bivariate normal scenarios with $\rho = -0.3$, while under the Cox model based scenario with $\beta = 1.25$ (equivalent to $\rho = -0.7$), the method is not able to deal with local perfect separation. Among all three methods, residual smoothing is the most stable one. Under bivariate normal based scenarios, the estimates show a relatively smaller bias than local Cox and plain Cox and the standard deviation is negligible, while under Cox model based scenarios, the performance of residual smoothing is more or less invariant over different censoring patterns and hazard shapes (Appendix B).

### 3.3.2 Non-parametric LOWESS and WMR

We choose {0.01, 0.33, 0.66} as the spans for LOEWSS, which indicates 1%, 33% and 66% of all intermediate estimates A($t$) would be considered when smoothing. For WMR estimator, the radii of the neighborhood (bandwidth) are chosen as {0.05, 0.35, 1.5}, where $n^{-0.2} = 0.35$ is the suggested bandwidth[10]. Thus, the lengths of the neighborhood are {0.1, 0.7, 3} in time scale. The percentages of intermediate A($t$) included in different size of neighborhood are shown in Table 3.1, for later time points, less A($t$)'s would be included. With increasing censoring percentages, the proportion of A($t$) included in the window may decrease since there are less events time points available.

Table 3.1: The corresponding percentages of intermediate estimates $A(t)$ included in WMR smoothing with certain bandwidth under bivariate normal scenario with 0% censoring (%).

| Time point | 0.135 | 0.233 | 0.368 | 0.607 | 1.000 | 1.649 | 2.718 | 4.482 | 7.389 |
|---|---|---|---|---|---|---|---|---|---|
| bandwidth=0.05 | 3.9 | 5.7 | 6.6 | 5.8 | 4.0 | 2.1 | 0.9 | 0.3 | 0.1 |
| bandwidth=0.35 | 23.5 | 28.9 | 37.0 | 39.5 | 28.5 | 15.3 | 6.3 | 2.0 | 0.5 |
| bandwidth=1.5 | 68.9 | 70.7 | 73.4 | 77.2 | 82.0 | 84.6 | 34.7 | 10.0 | 2.4 |

The LOWESS and WMR share the same intermediate concordance measure A($t$) defined as $\frac{\sum_{j\in R_0(t)} \mathbf{1}(M_j<M_i)+0.5\sum_{j\in R_0(t)} \mathbf{1}(M_j=M_i)}{n_0(t)}$. After a certain time point when no controls are available in the data ($n_0 = 0$), the intermediate concordance measure is not defined. The LOWESS estimated AUC($t$) is truncated, and there are no feasible estimates at $t = 7.39$, yet for WMR with bandwidth=1.5, the estimates are still available since there are still estimable intermediate results in this large neighborhood around $t = 7.39$ (Figure 3.9).

Accordingly, with the same intermediate estimates, their results could be almost overlapped under small span and bandwidth. For span = 0.01, the AUC($t$) curve for each simulated dataset would be very bumpy, and the standard deviation would be higher than WMR at the beginning (Figure 3.9). For WMR estimator, the smoothing window contains more events at earlier time points, which could smooth the curve in a good manner and reduce the variance. For later time points where events are extremely rare, the smoothing window might only contain one or two events, thus, the standard deviations converge to the results from LOWESS (Appendix A.1).

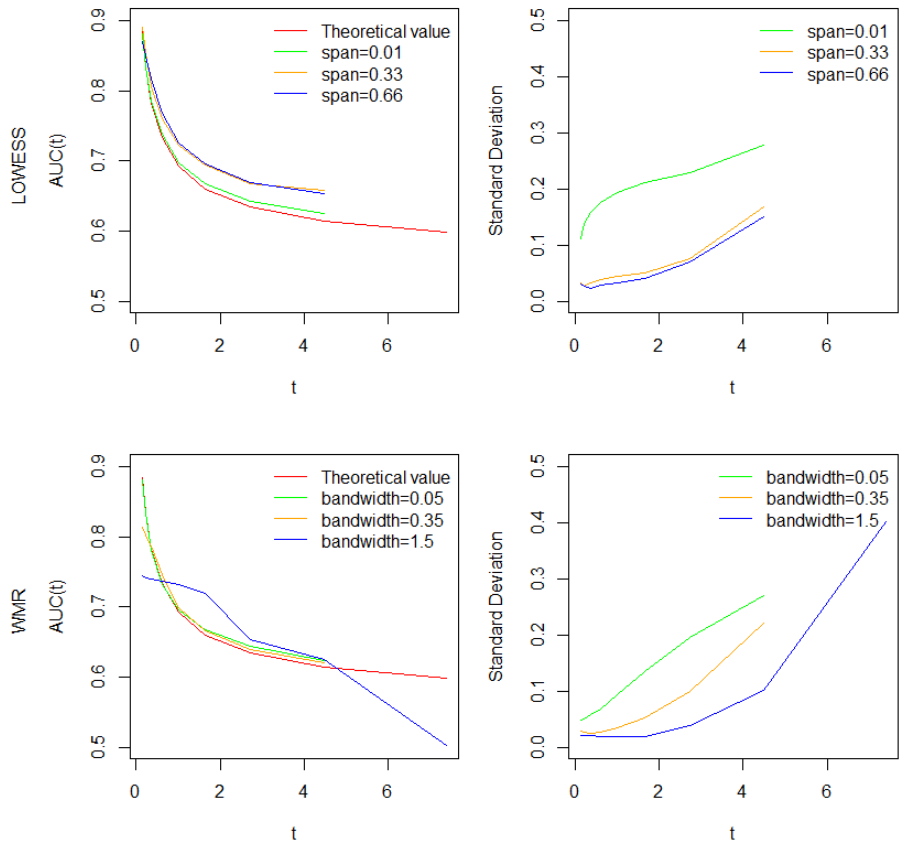


Figure 3.9: The performances of non-parametric LOWESS and WMR under bivariate normal scenario with 20% uniform censoring, $\rho = -0.7$. Upper panel: LOWESS. Bottom panel: WMR.

During the experiment we notice that both methods are sensitive to their span/bandwidth. With a larger span/bandwidth, the bias would increase, yet the standard deviation would decrease. Most of the events generated from $\mathrm{N}_2[0, 0, 1, 1, \rho]$ would happen in the beginning of follow-up, where the theoretical AUC values and the corresponding intermediate $\mathrm{A}(t)$ are relatively high. Our expectation is, with a large span/bandwidth, the estimates of later time points would pick up the earlier $\mathrm{A}(t)$ with high values, thus lead to the overestimation of $\mathrm{AUC}(t)$ and vise versa. From Figure 3.9, the LOWESS shows overestimation at later time points yet WMR underestimates the $\mathrm{AUC}(t)$ at earlier time points. This can be explained by the different mechanisms of including data points for smoothing. The LOWESS smoothing would first weight a certain proportion of data points around $t$ in terms of their distance among $t$-axis to obtained a prediction of $\mathrm{A}(t)$, then the residuals among $\mathrm{A}(t)$-axis would be calculated to adjust the original weights for making the final prediction [4]. In the beginning of data, most of the $\mathrm{A}(t)$ estimates are high, the LOWESS smoother is more likely to assign larger weights to these high $\mathrm{A}(t)$ values, and regards the lower $\mathrm{A}(t)$ values as outliers. Thus, for earlier time points, LOWESS would not underestimate the theoretical value as WMR did. For later time points, the LOWESS would pick up the high $\mathrm{A}(t)$'s from earlier time points, which may lead to overestimation. Yet the WMR smoother considers a moving window with certain radius around $t$ instead of the proportion, and the kernel weights are not adjusted by the $\mathrm{A}(t)$-axis residuals. For earlier time points, the lower $\mathrm{A}(t)$ values at later time points would be included and thus lead to underestimation. The trade-off has to be made during span/bandwidth selection. For WMR, the optimal bandwidth can be obtained by minimizing the integrated mean square error over time [10]. However, the

span selection method for LOWESS smoothing needs further investigation.

### 3.3.3 Semi-parametric Fractional polynomial

We implemented two fractional polynomial settings here: $p_1, ..., p_k = \{-2, -1, -0.5, 0, 0.5, 1, 2\}$ and $p_1, ..., p_k = \{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Figure 3.10 shows the results under the bivariate normal situation with 20% uniform censoring. The two settings behave similarly at the beginning, yet for later time points the fractional polynomial with degree = 9 performs slightly better. The simulation standard deviations under two settings do not differ much. Similar patterns are found under Cox model based scenarios (Figure 3.10, bottom panel). Under many scenarios (see Appendix A and B), the relatively large standard deviations for earlier time points indicates the fractional polynomial fit may be unstable in the beginning. For later time points, the standard deviations of both settings increase to 0.3 over time, which is not negligible compared to [0,1], the range of AUC($t$). For implementation, we adopted the function *optim* in R as suggested by the authors. However, this function can be trapped by local minima and thus influence the final estimation, which may need further investigation.

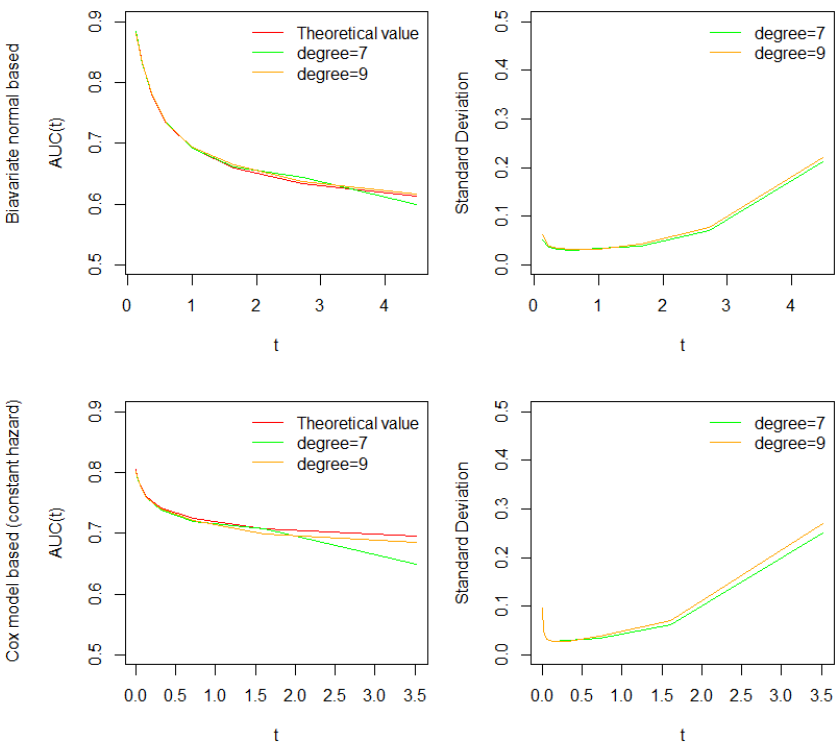


Figure 3.10: The performances of semi-parametric estimates from fractional polynomials under 20% uniform censoring. Upper panel: bivariate normal scenario with $\beta = -0.7$. Bottom panel: Cox model based scenario, $\beta = 1.25$.

### 3.3.4 Computation time

From Table 3.2, the two non-parametric methods are the fastest among all methods. For Cox model based methods, local Cox is the most time-consuming one, while residual smoothing is slightly faster and plain Cox is least time-consuming. For fractional polynomials, the optimization procedure under function *optim* takes a longer time, thus with a higher degree polynomial, the computation time would increase. Note that the optimal bandwidth selection for non-parametric LOWESS, WMR, local Cox and residual smoothing could be computationally expensive [12], which is not included in this study. The plain Cox and fractional polynomials do not require such optimization procedure. Thus Table 3.2 only offers an overview of the recorded computation time of this study, for more comparable computation time, we need further investigation.

Table 3.2: The computation time of the methods under default setting. The time is recorded under bivariate normal scenario, $\rho = -0.3$ for all censoring patterns (1000 replications each).

| Methods | Computation time (minute) |
|---|---|
| Plain Cox | 18.22 |
| Local Cox | 2148.67 |
| Residual Smoothing | 75.33 |
| Non-parametric LOWESS and WMR | 8.25 |
| Fractional polynomial (degree = 7) | 1111.37 |
| Fractional polynomial (degree = 9) | 1872.52 |

## 3.4 Overall comparison

Under bivariate normal scenarios, $\rho = -0.7$ (Appendix A.1), the performances of local Cox and residual smoothing are satisfactory for data with small censoring percentages (i.e. 20%). However, for data with more than 20% censoring, these two methods are not able to retain a good fit for later time points. The proportional hazards assumption is not satisfied under these scenarios, thus plain Cox shows poor performance with large bias. The non-parametric LOWESS and WMR cannot estimate the AUC($t$) when there are no available controls in the data. Nevertheless, these methods still pick up the trend of AUC($t$) for data with up to 60% censoring. For the trade-off between bias and simulation standard deviations, an appropriate span (bandwidth) needs to be specified either by bootstrapping or other optimization methods. The fractional polynomial with suggested degree and the fractional polynomial with two more terms perform similar in terms of bias and standard deviations. The sizable standard deviations for earlier and later time points need to be taken into account before applying this method. However, adding fractional polynomial terms will significantly increase the computation time, consider the negligible gain, it may not be a good choice. Under bivariate normal scenarios, $\rho = -0.3$ (Appendix A.2), the Cox model based methods, especially local Cox, show irregular and considerable bias. The non-parametric methods also show some extreme estimates (i.e. AUC($t$) = 1) at later time points. The fractional polynomials are relatively stable in terms of bias, yet the simulation standard deviations, as mentioned before, are non-negligible. For the Cox model based scenarios, the plain Cox shows an evident improvement since the proportional hazards assumption is satisfied. When $\beta = 1.25$, the local Cox can be hindered by the perfect separation problem within the suggested span ($n^{-0.2} = 0.35$). When $\beta = 0.4$, the non-parametric

WMR (span=0.35) surpasses other methods in terms of bias and simulation standard deviation.

In general, for earlier time points, the performance between different methods (except LOWESS) under default settings are fairly satisfactory, while for later time points, all methods are not able to retain a good fit since the events rarely occur. Although the non parametric methods and the fractional polynomial show relatively small bias, their standard deviations are non-negligible. When the marker has a mediocre effect on event times, the local Cox and residual smoothing, with smaller bias and standard deviations, are relatively stable. When the effect size is small, the non-parametric WMR and LOWESS, and fractional polynomials show less biased results yet with large standard deviations, note that the non-parametric WMR (span=0.35) has comparable standard deviations with Cox model based methods under some scenarios.

# Chapter 4

# Trend of I/D AUC(t)

In previous chapters, we noticed that the I/D AUC values under both Bivariate normal and Cox model based settings have a decreasing trend over time. The intuitive expectation for this decreasing trend is that, under the Cox proportional hazards model, the group of patients becomes more homogeneous over time since the patients with larger marker values will drop out earlier, and with a more homogeneous group, the concordance measure will be lower. Thus, the I/D AUC would be expected to be a decreasing function of $t$.

Yet Pepe et al [7] stated that the I/D AUC could be an increasing function of time. For this statement, we consider a scenario with a binary marker defined as 1 for subjects who experience the event and 0 for event-free subjects. We assume that the event times are untied, and the censoring happens after the last event time. For I/D AUC, the case set and control set vary over time. Consider the extreme case, at each time point, the cases would not be predicted as controls. Thus, in Table 1.1, we have $c = 0$ and the true positive rate at $t_0$ equals 1. For the false positive rate, the b is the number of subjects who will have an event before the end of the follow-up, and d is the number of subjects survive until the end of the follow-up. Note that d is a constant overtime in this case. Thus, the false positive rate at $t_0$ is $\frac{b}{b+d}$.

Table 4.1: Cross table at time $t_0$. Under this scenario, $c = 0$. The size of case set is 1, and the size of control set is $b + d$. The incident true positive rate is $\frac{1}{1+c} = 1$, and the dynamic false positive rate is $\frac{b}{b+d}$.

| | Observed case | Observed control |
|---|---|---|
| Predicted case | 1 | $b$ |
| Predicted control | $c$ | $d$ |

At later time points, we would have a decreasing false positive rate due to the control set dropped subjects with higher marker values (in this case, 1) who had events. Table 4.1 shows the cross table at a later time point $t_1$. The dynamic false positive rate is $\frac{b-1}{b+d-1}$, which is smaller than that of $t_0$ ($\frac{b}{b+d}$). Thus, with a constant true positive rate, the ROC curve would shift to the left from $t_0$ to $t_1$, and the I/D AUC would be an increasing function of time[7].

Table 4.2: Cross table at time $t_1$. Under this scenario, $c = 0$. The size of case set is 1, and the size of control set is $b + d - 1$. The incident true positive rate is $\frac{1}{1+c} = 1$, and the dynamic false positive rate is $\frac{b-1}{b+d-1}$.

| | Observed case | Observed control |
|---|---|---|
| Predicted case | 1 | $b-1$ |
| Predicted control | $c$ | $d$ |

However, the previous researchers did not provide a rigorous proof of the monotonicity of I/D AUC. In this chapter, we evaluate the trend of I/D AUC on a theoretical level by taking the first order derivative of its expression.

## 4.1 Trend of I/D AUC under the Cox model with a binary marker

In this section, we consider a case that the event times follow a Cox model with a baseline hazard $(\Lambda_0(t), \lambda_0(t))$ and a binary marker $X$, with $P(X = 1) = p$ and $P(X = 0) = 1-p$. The Cox model is defined as $\lambda(t \mid X = x) = \lambda_0(t)\exp(\beta X)$, where $\beta$ is the effect size. The survival function and density function of this model can be written as $S(t \mid X = x) = \exp[-\Lambda_0(t) \cdot \exp(\beta X)]$ and $f(t \mid X = x) = \lambda(t \mid X = x)S(t \mid X = x) = \lambda_0(t)\exp[\beta X - \Lambda_0(t) \cdot \exp(\beta X)]$. Here we only show the situation where the marker and event times are negatively correlated, i.e. $\exp(\beta) > 1$, since the results obtained under a corresponding positive correlation is symmetrical to the results from the negative correlation with respect to the line $\mathrm{AUC}(t) = 0.5$.

We have $\lambda(t \mid X = 0) = \lambda_0(t)$. For simplicity, we denote $\lambda(t \mid X = 1)$ as $\lambda_1(t)$, $S(t \mid X = 0)$ as $S_0(t)$, $S(t \mid X = 1)$ as $S_1(t)$, $f(t \mid X = 0)$ as $f_0(t)$, $f(t \mid X = 1)$ as $f_1(t)$, and the hazard ratio $\exp(\beta)$ as $\eta$.

As mentioned in Chapter 2, the $\mathrm{TP}^I(t)$ and $\mathrm{FP}^D(t)$ can be written as (4.1) and (4.2), where the marginal density $f_T(t)$ and the survival functions $S_T(t)$ of $t$ are defined as $f_T(t) = \sum_x f(t, x) = f_1(t)p + f_0(t)(1-p)$ and $S_T(t) = \sum_x S(t, x) = S_1(t)p + S_0(t)(1-p)$, and $c$ is the cut-off point of marker $X$.

$$\mathrm{TP}^I(t) = \mathrm{sensitivity}^I(c, t) = P(X_i > c \mid T_i = t) = \frac{\sum_{x>c} f(t \mid X = x)P(X = x)}{f_T(t)} \tag{4.1}$$

$$\mathrm{FP}^D(t) = 1 - \mathrm{specificity}^D(c, t) = P(X_i > c \mid T_i > t) = \frac{\sum_{x>c} S(t \mid X = x)P(X = x)}{S_T(t)} \tag{4.2}$$

According to these definitions, when $C = 1$, we have $\mathrm{TP}^I(t) = \mathrm{FP}^D(t) = 0$, while when $C = 0$, $\mathrm{TP}^I(t) = \mathrm{FP}^D(t) = 1$. Equations (4.3) and (4.4) show the $\mathrm{TP}^I(t)$ and $\mathrm{FP}^D(t)$ rate when $C \in (0, 1)$.

$$\mathrm{TP}^I(t) = \frac{f_1(t)p}{f_1(t)p + f_0(t)(1-p)} = \frac{\lambda_1(t)S_1(t)p}{\lambda_1(t)S_1(t)p + \lambda_0(t)S_0(t)(1-p)} \tag{4.3}$$

$$\mathrm{FP}^D(t) = \frac{S_1(t)p}{S_1(t)p + S_0(t)(1-p)} \tag{4.4}$$

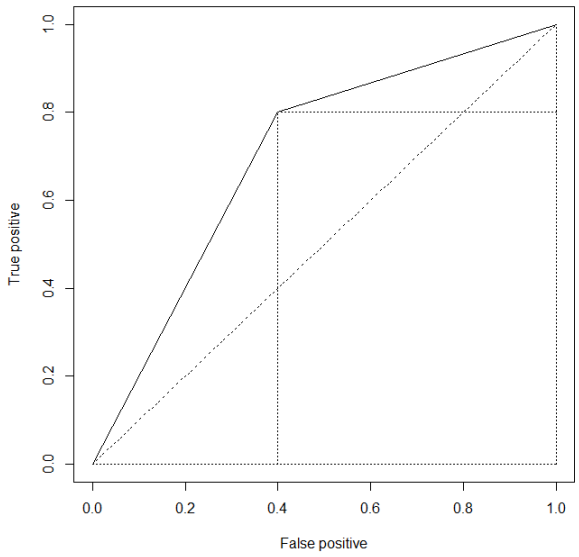


Figure 4.1: The shape of the ROC curve with one binary marker. The AUC can be divided into two triangles and one rectangular.

As shown in Figure 4.1, the AUC($t$) consists of two triangles and one rectangular, thus it can be simplified as follows.

$$\begin{aligned}\mathrm{AUC}(t) &= 0.5 \cdot \mathrm{TP}^I(t) \cdot \mathrm{FP}^D(t) + 0.5 \cdot [1 - \mathrm{TP}^I(t)] \cdot [1 - \mathrm{FP}^D(t)] + [1 - \mathrm{FP}^D(t)] \cdot \mathrm{TP}^I(t) \\ &= 0.5 \cdot [\mathrm{TP}^I(t) - \mathrm{FP}^D(t)] + 0.5.\end{aligned} \tag{4.5}$$

To evaluate the trend of AUC($t$), we take its first order derivative of (4.5) as follows.

$$\begin{aligned}\frac{\partial \mathrm{AUC}(t)}{\partial t} &= \frac{\partial\{0.5[\mathrm{TP}^I(t) - \mathrm{FP}^D(t)] + 0.5\}}{\partial t} \\ &= 0.5 \cdot \frac{\partial\{[\mathrm{TP}^I(t) - \mathrm{FP}^D(t)]\}}{\partial t}\end{aligned} \tag{4.6}$$

We can rewrite $\mathrm{TP}^I(t) - \mathrm{FP}^D(t)$ as

$$\begin{aligned}\mathrm{TP}^I(t) - \mathrm{FP}^D(t) &= \frac{\lambda_1(t)S_1(t)p \cdot [S_1(t)p + S_0(t)(1-p)] - S_1(t)p \cdot [\lambda_1(t)S_1(t)p + \lambda_0(t)S_0(t)(1-p)]}{[\lambda_1(t)S_1(t)p + \lambda_0(t)S_0(t)(1-p)] \cdot [S_1(t)p + S_0(t)(1-p)]} \\ &= \frac{S_0(t)S_1(t)p(1-p)[\lambda_1(t) - \lambda_0(t)]}{[\lambda_1(t)S_1(t)p + \lambda_0(t)S_0(t)(1-p)] \cdot [S_1(t)p + S_0(t)(1-p)]} \\ &= \frac{S_0(t)S_1(t)p(1-p)(\eta - 1)}{[\eta S_1(t)p + S_0(t)(1-p)] \cdot [S_1(t)p + S_0(t)(1-p)]}.\end{aligned} \tag{4.7}$$

In order to obtain the first order derivative of (4.7), we separately calculate the derivatives of the numerator and the denominator. We denote the numerator of (4.7) as $g(t)$ and the denominator of (4.7) as $d(t)$.

$$\begin{aligned}\frac{\partial g(t)}{\partial t} &= -\left(\eta^2-1\right)\cdot\lambda_0(t)\cdot p\cdot(1-p)\cdot S_0(t)\cdot S_1(t)\\ \frac{\partial d(t)}{\partial t} &= -\lambda_0(t)\{[\eta^2\cdot p\cdot S_1(t)+(1-p)\cdot S_0(t)][p\cdot S_1(t)+(1-p)\cdot S_0(t)]\\ &\quad +[\eta\cdot p\cdot S_1(t)+(1-p)\cdot S_0(t)]^2\}\end{aligned} \tag{4.8}$$

To further simplify the equation, we evaluate (4.6) at $t=0$, thus $S_1(0)=S_0(0)=1$ and we will write $\lambda_0(0)$ as $\lambda_0$.

$$\begin{aligned}\frac{\partial\{[\mathrm{TP}^I(t)-\mathrm{FP}^D(t)]\}}{\partial t}\Big|_{t=0} &= \frac{\frac{\partial g(t)}{\partial t}\cdot d(t)-\frac{\partial d(t)}{\partial t}\cdot g(t)}{d(t)^2}\Big|_{t=0}\\ &= \frac{(\eta-1)^2\cdot p\cdot(1-p)\cdot\lambda_0[(\eta-1)\cdot p^2+2p-1]}{[\eta\cdot p+(1-p)]^2}\end{aligned} \tag{4.9}$$

The denominator of equation (4.9) is non-negative. It can also be observed that $(\eta-1)^2\cdot p\cdot(1-p)\cdot\lambda_0$ is non-negative, and thus sign of (4.9) depends on $(\eta-1)\cdot p^2+2p-1$a parabola of $p$. The determinant of this parabola equals $\Delta=4+4\cdot(\eta-1)$. For a given hazard ratio $\eta$, the solution to $(\eta-1)\cdot p^2+2p-1$ is $p=(\sqrt{\eta}+1)^{-1}$, restricted to $p\in[0,1]$ and $\eta\geq 1$. We already know that when $\eta=1$, $AUC(T)\equiv 0.5$. For a given $\eta>1$, when $p\in\left[0,(\sqrt{\eta}+1)^{-1}\right)$, (4.9) is positive, while for $p\in\left((\sqrt{\eta}+1)^{-1},1\right]$, (4.9) is negative. Figure 4.2 shows the contour plot of (4.9) under $\lambda_0=2$. There are some positive derivatives at $t=0$, which indicate there are increasing trends.

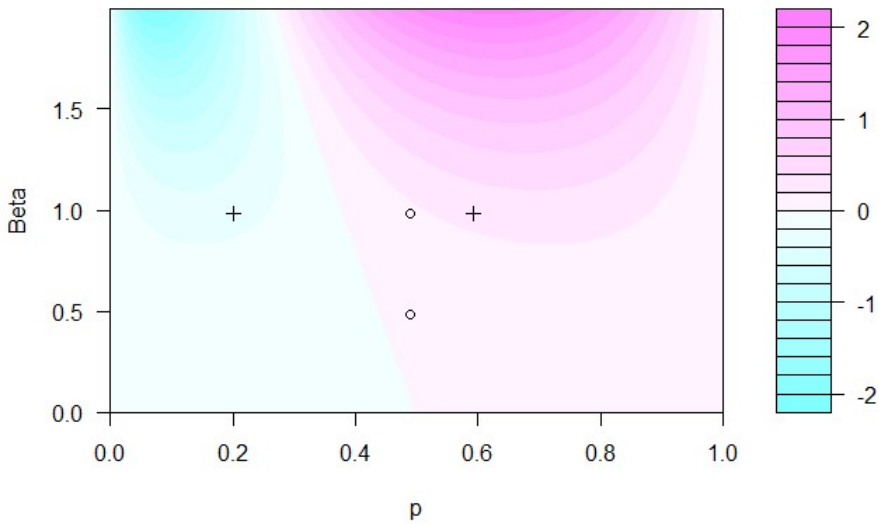


Figure 4.2: The contour plot of $\frac{\partial\{[\mathrm{TP}^I(t)-\mathrm{FP}^D(t)]\}}{\partial t}$ at $t=0$, $\lambda_0=2$.

We choose the Cox models with different parameters for illustration. The theoretical AUC values of the Cox models are calculated by (4.5) and (4.7). We first consider common scenarios where $p = 0.5$. It can be observed in Figure 4.3 that even with balanced marker values, the I/D AUC($t$) may show non-monotonic trends over time.

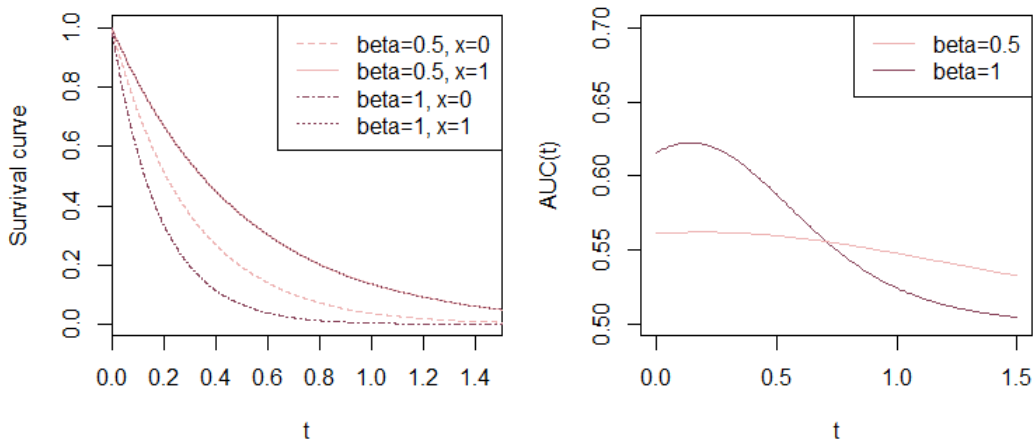


Figure 4.3: Left panel: survival curve of the Cox models with $p = 0.5$, $\beta = 1$ and $\beta = 0.5$ as indicated by 'o' in Figure 4.2, the curves of $x$=0 are overlapped. Right panel: the theoretical AUC(t) values of the Cox models.

On the other hand, we consider two scenarios with $p$ varied under $\beta = 1$. According to Figure 4.4, under $p = 0.6$, we observe that the AUC($t$) first increases then decreases over time, while under $p = 0.2$, the AUC($t$) shows a decreasing trend over time. Thus, under the Cox model with binary marker, the I/D AUC is not always a decreasing function of time.

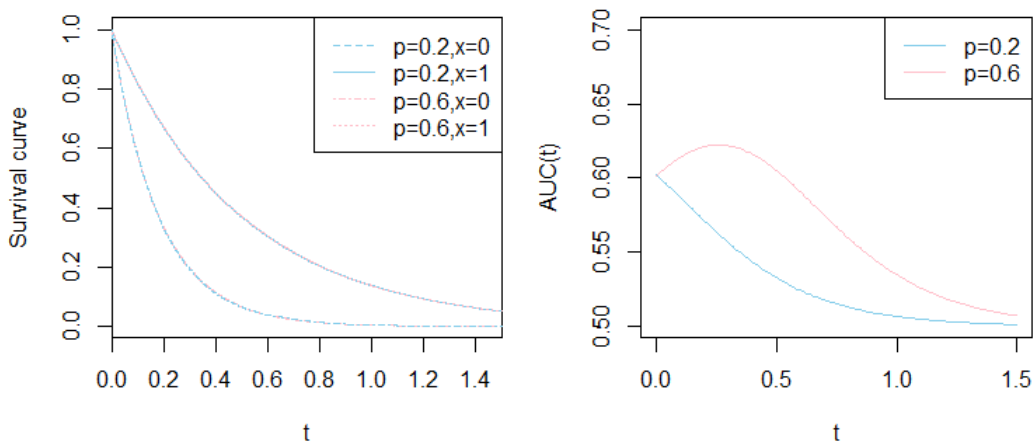


Figure 4.4: Left panel: survival curve of the Cox models with $\beta = 1$, $p = 0.2$ and $p = 0.6$ as indicated by '+' in Figure 4.2. Here the two models have the same $\beta$'s, thus the survival curves overlapped. Right panel: the theoretical AUC(t) values of the Cox models.

## 4.2 Simulation based evaluation

In this section, we simulate 1000 datasets with a larger sample size 1000 from the two aforementioned Cox models ($\beta = 1$, $p$=0.2 and 0.6). There is no censoring in the datasets. We would like to know whether the methods, namely Plain Cox, residual smoothing, non-parametric LOWESS and WMR, and fractional polynomial are able to pick up the trend of I/D AUC under these scenarios. Here local Cox is not included because of the limitation of the computation time. The span for residual smoothing and the bandwidth of WMR equal $1000^{-0.2} = 0.25$, and the span for LOWESS is 0.66 by default.

Figure 4.5 shows the results of the simulation study. When $p = 0.6$, the theoretical and estimated AUC values show a non-monotonic trend over time, while when $p = 0.2$, they show a decreasing trend over time. It can be observed that plain Cox, residual smoothing show very small bias and simulation standard deviation, which matches our expectation since the datasets are generated from Cox models. The fractional polynomial with 7 degrees also shows small bias but with larger standard deviations. For the non-parametric methods, WMR shows larger bias since the suggested bandwidth 0.25 may be too wide under these scenarios. The non-parametric LOWESS fails to capture the trend of true values under both scenarios, especially for the second scenario, the estimates are below 0.5. This again shows the default bandwidth may not be a permissible choice when adopting LOWESS smoothing for I/D AUC estimates.

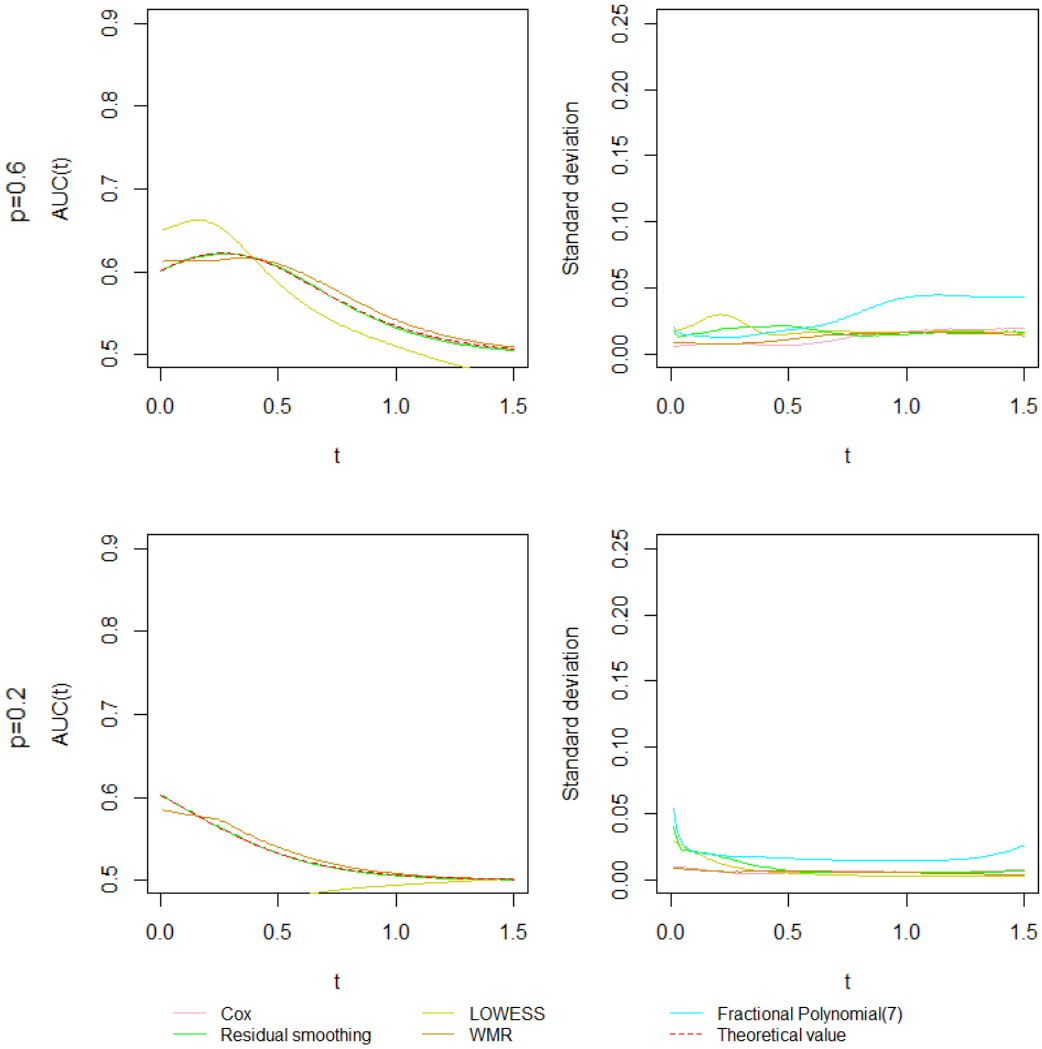


Figure 4.5: The performances under the Cox model with a baseline hazard $\lambda_0(t) = 2$, 0% censoring, $\beta = 1$. Upper panel: $p = 0.6$. Bottom panel: $p = 0.2$.

# Chapter 5

# Discussion

## 5.1 Different estimation methods of I/D AUC

The first part of this thesis compared several methods of estimating I/D AUC under a series of simulation studies. The methods proposed by Heagerty and Zheng [6] based on the Cox model were more stable than other methods in terms of the standard deviations under most of the scenarios. However, when the effect size of the marker is small, the Cox model based methods are not recommended. The plain Cox showed better performance only when the proportional hazards assumption was satisfied. The residual smoothing and the local Cox have a wider range of use since they do not require the proportional hazards assumption. Comparing these two methods, the local Cox was not feasible with a large effect size, while residual smoothing showed more consistent results. Thus, among the Cox model based methods, the residual smoothing is recommended when the marker has a mediocre to strong effect. For the non-parametric LOWESS, the default span would lead to overestimation and over smoothing. For WMR, a non-proper bandwidth would also lead to large bias of the estimates. The trade-off between span/bandwidth and the robustness of the method should be made. For the fractional polynomial, the gain of adding extra polynomial terms was not pronounced. Different censoring patterns did not influence the performance of the methods much.

In this study, most of the methods adopt kernel smoothing. Under this circumstance, varying the shape and the span of the kernel can be an option to enhance their performance. For non-parametric WMR and LOWESS, the performance of final estimates hinges on the span (bandwidth) we choose. LOWESS weights a certain percentage (span) of the data points by their x-axis distances and adjusts the weights by y-axis residuals, while WMR incorporate the data points in the one-dimensional neighborhood defined by the bandwidth. The optimal bandwidth of WMR can be obtained by optimizing the asymptotic integrated mean square error [10], which still needs to be implemented. While for LOWESS smoothing, a common way to determine the span is K-fold cross-validation. As shown in chapter 3, an ideal fit over time is hard to obtain under a certain span or bandwidth. We could also consider the adaptive span or bandwidth over time, however, this alternative may be more computationally intensive. For local Cox and residual smoothing, the smooth kernel used in package *risksetROC* is the Epanechnikov kernel, while the non-parametric WMR uses the uniform kernel (arithmetic mean). Nonetheless, other kernels may also be adopted to vary the weights of the point estimates. For instance, the

Gaussian kernel assigns more weight on the two-sided tail part of the neighborhood compared to Epanechnikov kernel, which may retain more information from previous and future time points. For fractional polynomial, it does not require bandwidth or kernel selection. An improvement of this method may be to speed up the computation and do not get trapped in local minima by using the function *glm* instead of the function *optim* offered in the supplementary code [12].

## 5.2 Trend of I/D AUC

Our expectation was that the I/D AUC would be a decreasing function of time. In the simulation studies, we observed that the theoretical I/D AUC was decreasing over time under bivariate normal scenarios and under Cox model based scenarios. In this second part, for simplification, we considered the Cox model with a binary marker and evaluated the partial derivative of I/D AUC($t$) according to $t$ at time 0. Surprisingly, with certain parameters, the I/D AUC showed a non-monotonic trend over time. The intuition behind is that false positive rate declines faster than the true positive rate at certain time range. However, the reason behind the different decreasing speeds of true and false positive rates is not clear.

For further investigation, we simulated datasets from Cox models with different trend of I/D AUC to see whether the aforementioned methods were able to pick up the trend. The simulation study showed that the non-monotonic trend of I/D AUC did exist. The plain Cox, residual smoothing and fractional polynomial perfectly picked up the non-monotonic trend of I/D AUC. The satisfactory performance of Cox model based methods is not surprising since the datasets are generated from Cox models, while the good performance of the fractional polynomial shows its flexibility to accommodate different shapes of I/D AUC curves. The non-parametric WMR was biased since the suggested span was not proper, and the same problem was found for non-parametric LOWESS. Furthermore, for imbalanced binary marker values, non-parametric LOWESS with default span could have misleading results.

We analyzed the Cox model with binary markers in this study. For bivariate normal scenarios, the explicit expression of the first order derivative of I/D AUC is difficult to obtain. Yet other continuous markers (i.e. marker generated from uniform distribution or other continuous distributions) could also be considered. In addition, we evaluate the partial derivative at time 0. However, for a detailed explanation for the non-monotonic trend of I/D AUC, the derivative of whole time range should also be studied. Future studies should focus on these aspects.

# Bibliography


[1] Paul Blanche, Aurélien Latouche, and Vivian Viallon. Time-dependent AUC with right-censored data: a survey. In *Risk assessment and evaluation of predictions*, pages 239–251. Springer, 2013.

[2] Zongwu Cai and Yanqing Sun. Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics*, 30(1):93–111, 2003.

[3] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[4] William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

[5] Patricia M Grambsch and Terry M Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

[6] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005.

[7] Margaret S Pepe, Yingye Zheng, Yuying Jin, Ying Huang, Chirag R Parikh, and Wayne C Levy. Evaluating the ROC performance of markers for future events. *Lifetime data analysis*, 14(1):86–113, 2008.

[8] Dimitris Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 2011.

[9] Patrick Royston and Douglas G Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied statistics*, pages 429–467, 1994.

[10] P Saha-Chaudhuri and PJ Heagerty. Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics*, 14(1):42–59, 2013.

[11] Matthias Schmid and Sergej Potapov. A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in medicine*, 31(23):2588–2609, 2012.

[12] Weining Shen, Jing Ning, and Ying Yuan. A direct method to evaluate the time-dependent predictive accuracy for biomarkers. *Biometrics*, 71(2):439–449, 2015.

[13] Xiao Song and Xiao-Hua Zhou. A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistica Sinica*, pages 947–965, 2008.

[14] Hans van Houwelingen and Hein Putter. *Dynamic Prediction in Clinical Survival Analysis*, chapter 3. CRC Press, 2011.

[15] Ronghui Xu and John OQuigley. Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1(4):423–439, 2000.

# Appendix A

## Simulation results: bivariate normal based scenarios

### A.1 $\rho = -0.7$

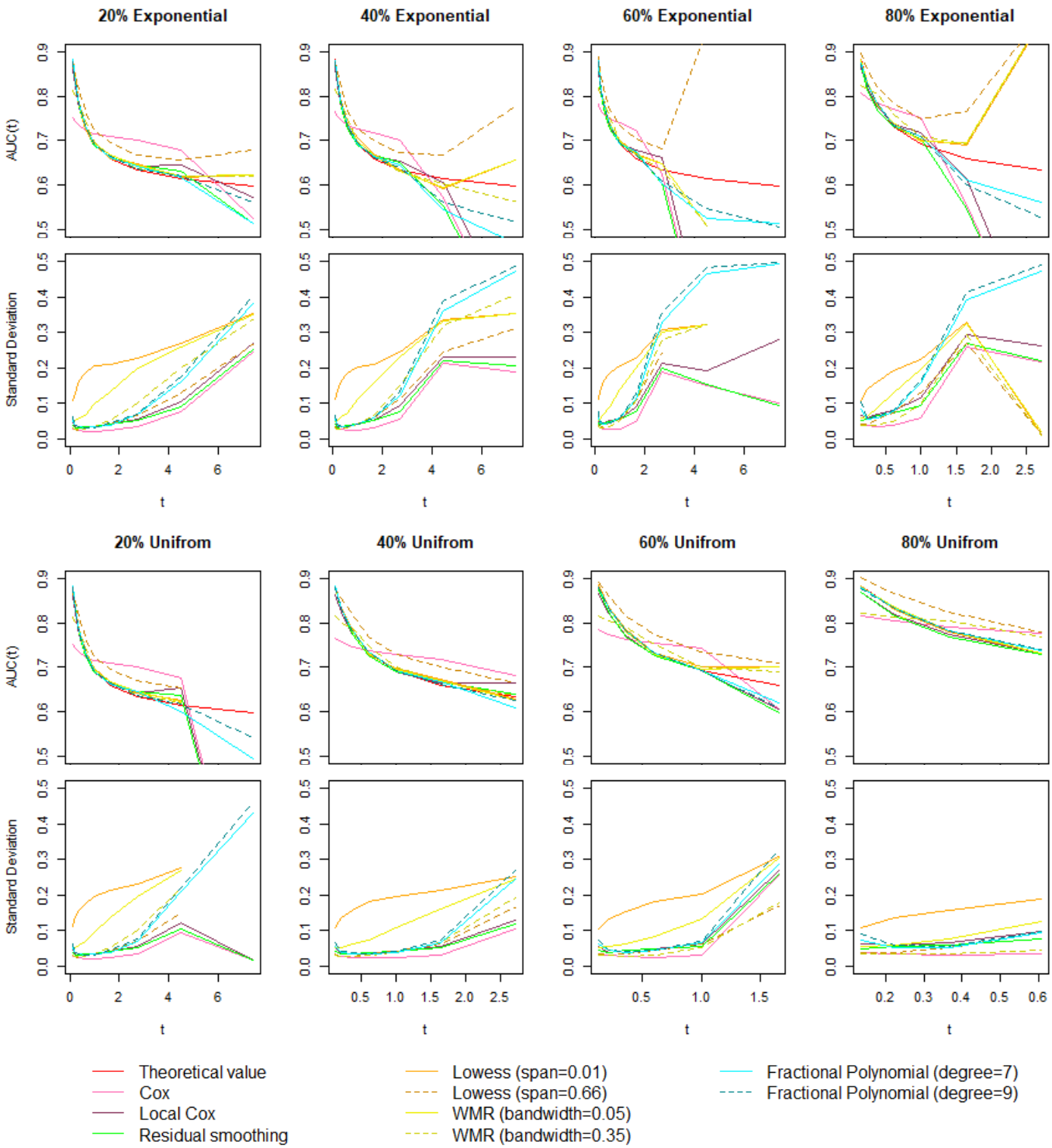


Figure A.1: Simulation scenario: bivariate normal with $\rho = -0.7$.

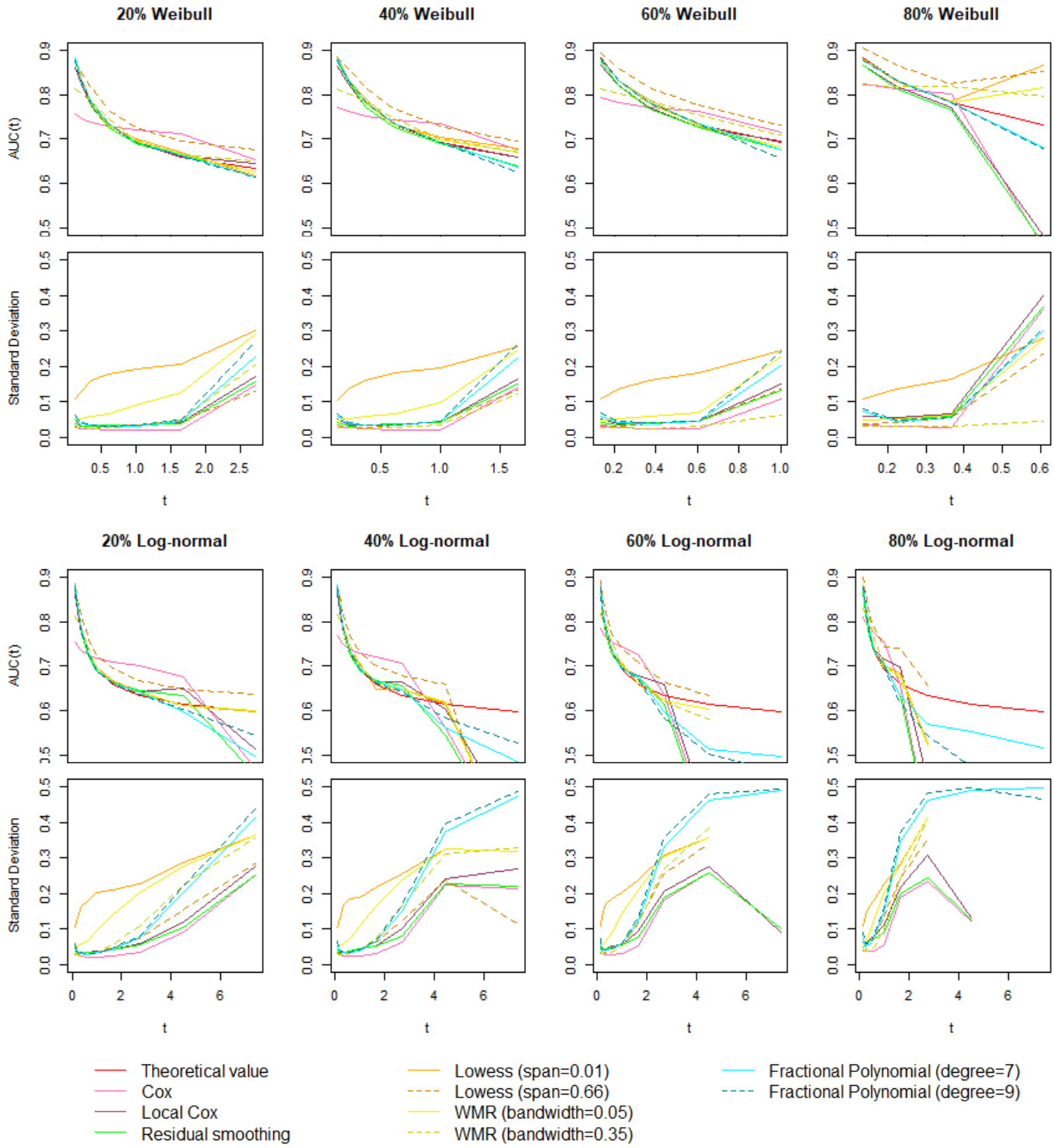


Figure A.2: Simulation scenario: bivariate normal with $\rho = -0.7$.

## A.2 $\rho = -0.3$

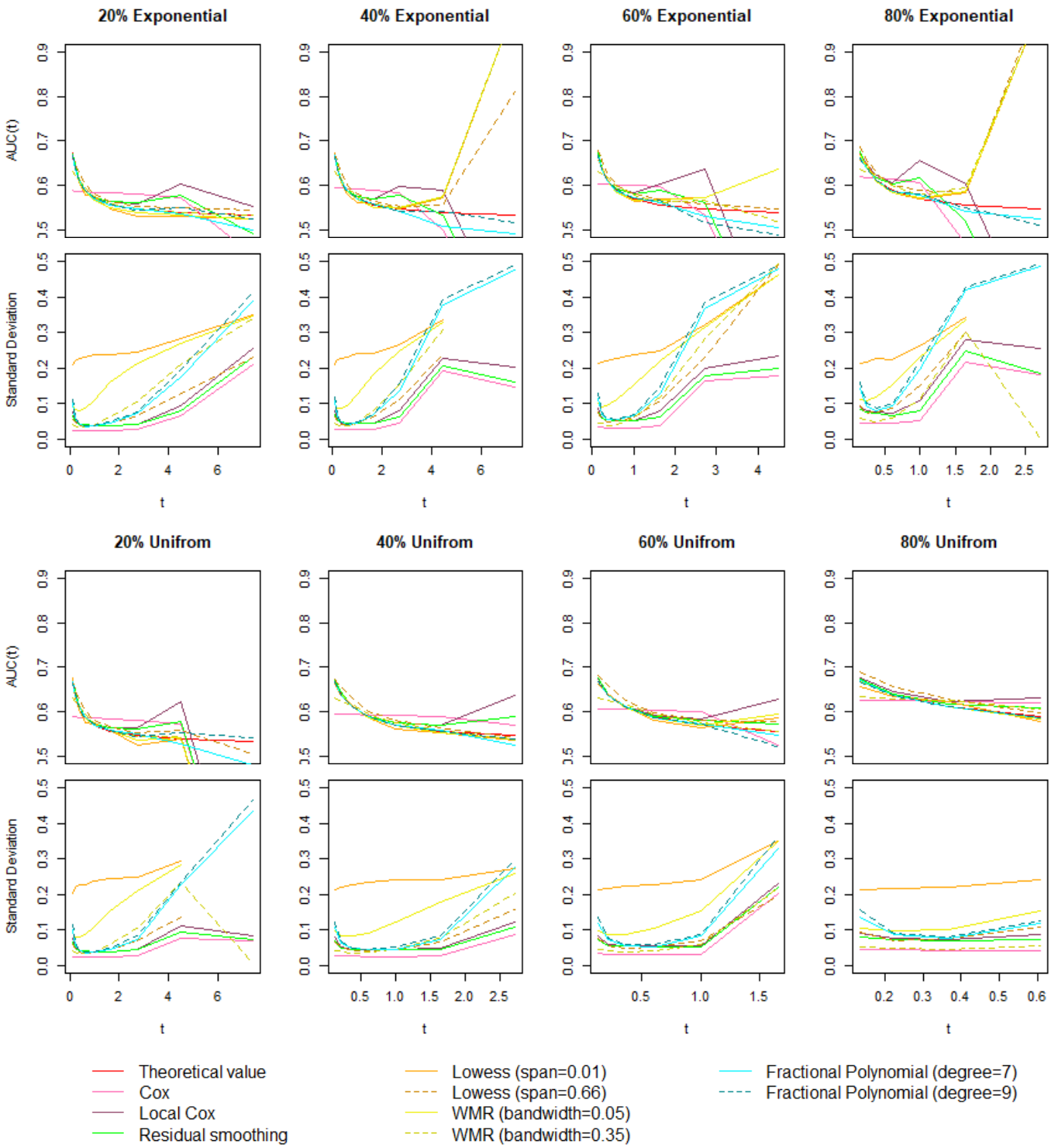


Figure A.3: Simulation scenario: bivariate normal with $\rho = -0.3$.

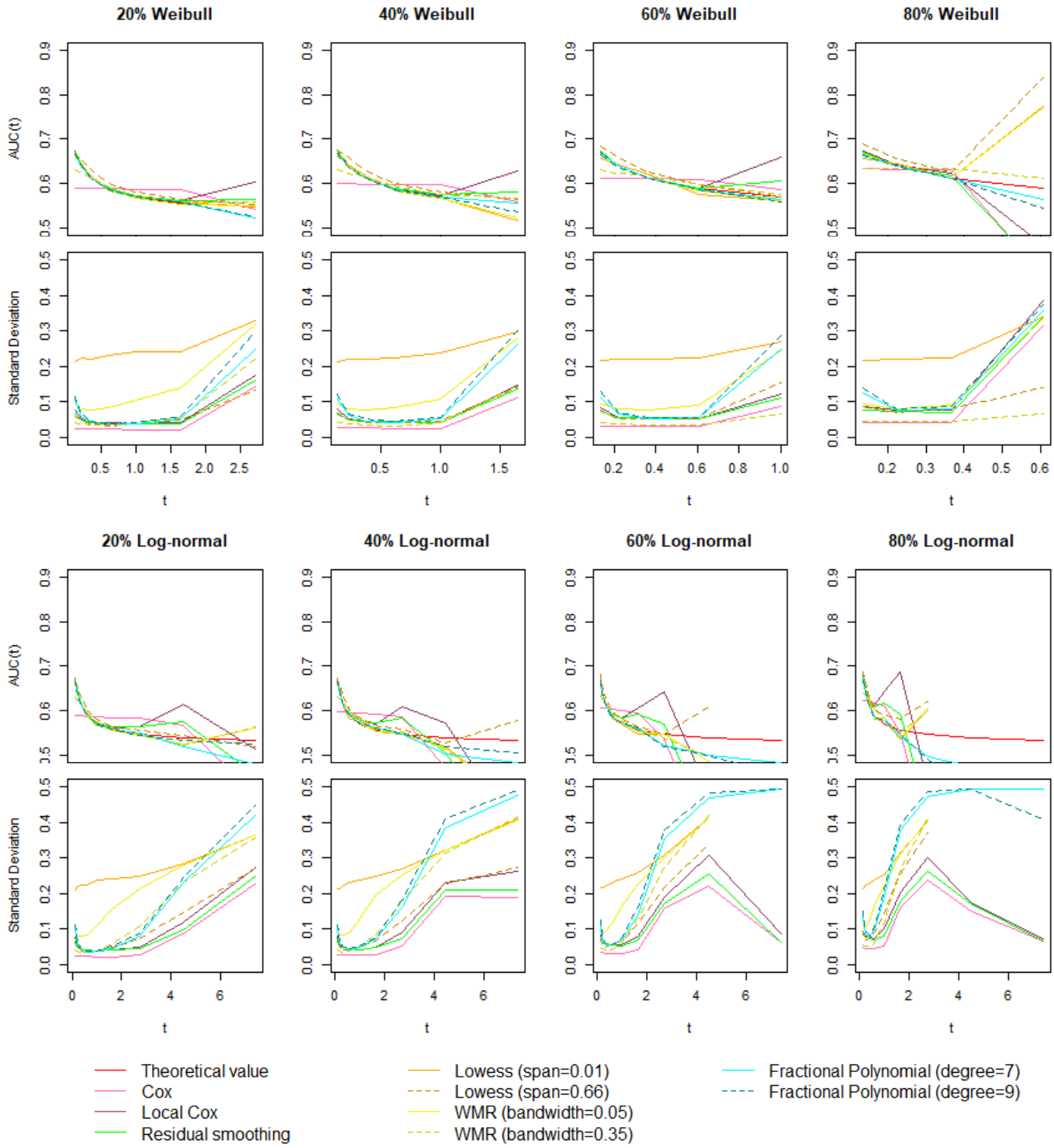


Figure A.4: Simulation scenario: bivariate normal with $\rho = -0.3$.

# Appendix B

## Simulation results: Cox model based scenarios

### B.1 $\beta = 1.25$

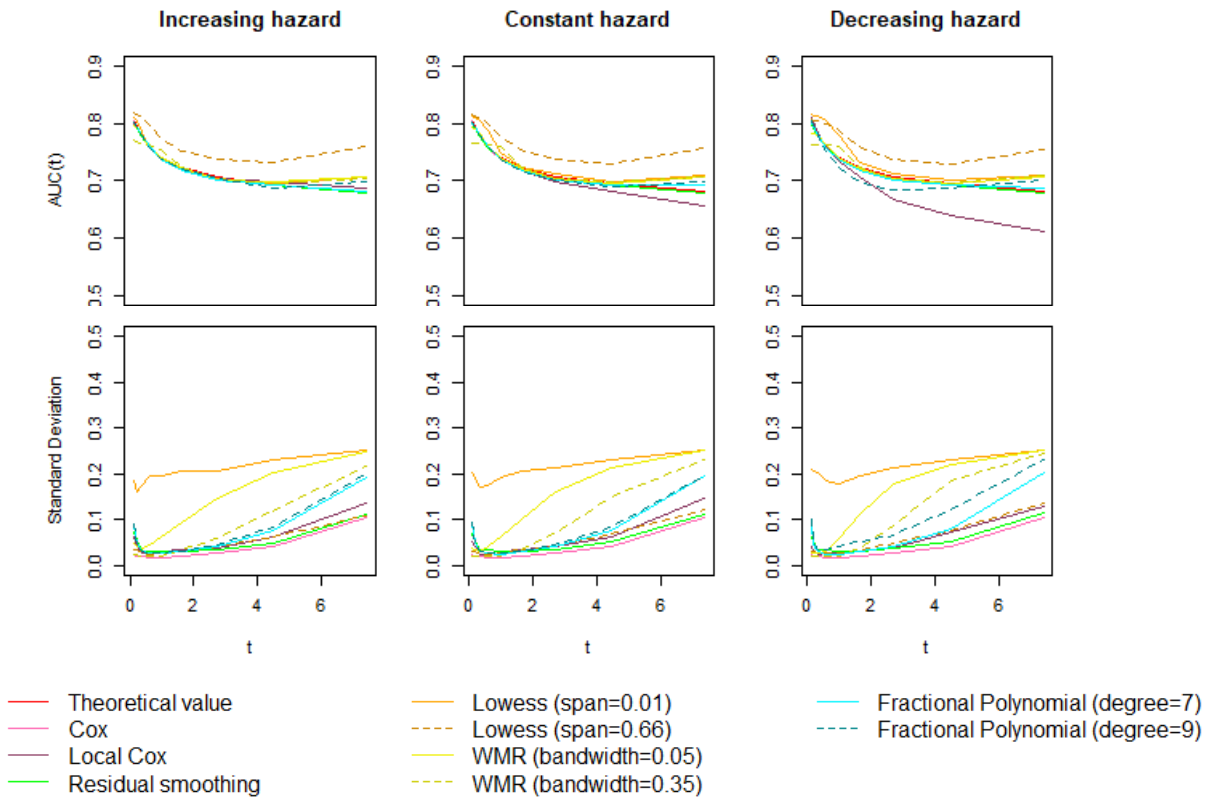


Figure B.1: Simulation scenario: Cox model based data with $\beta = 1.25$, 0% censoring.

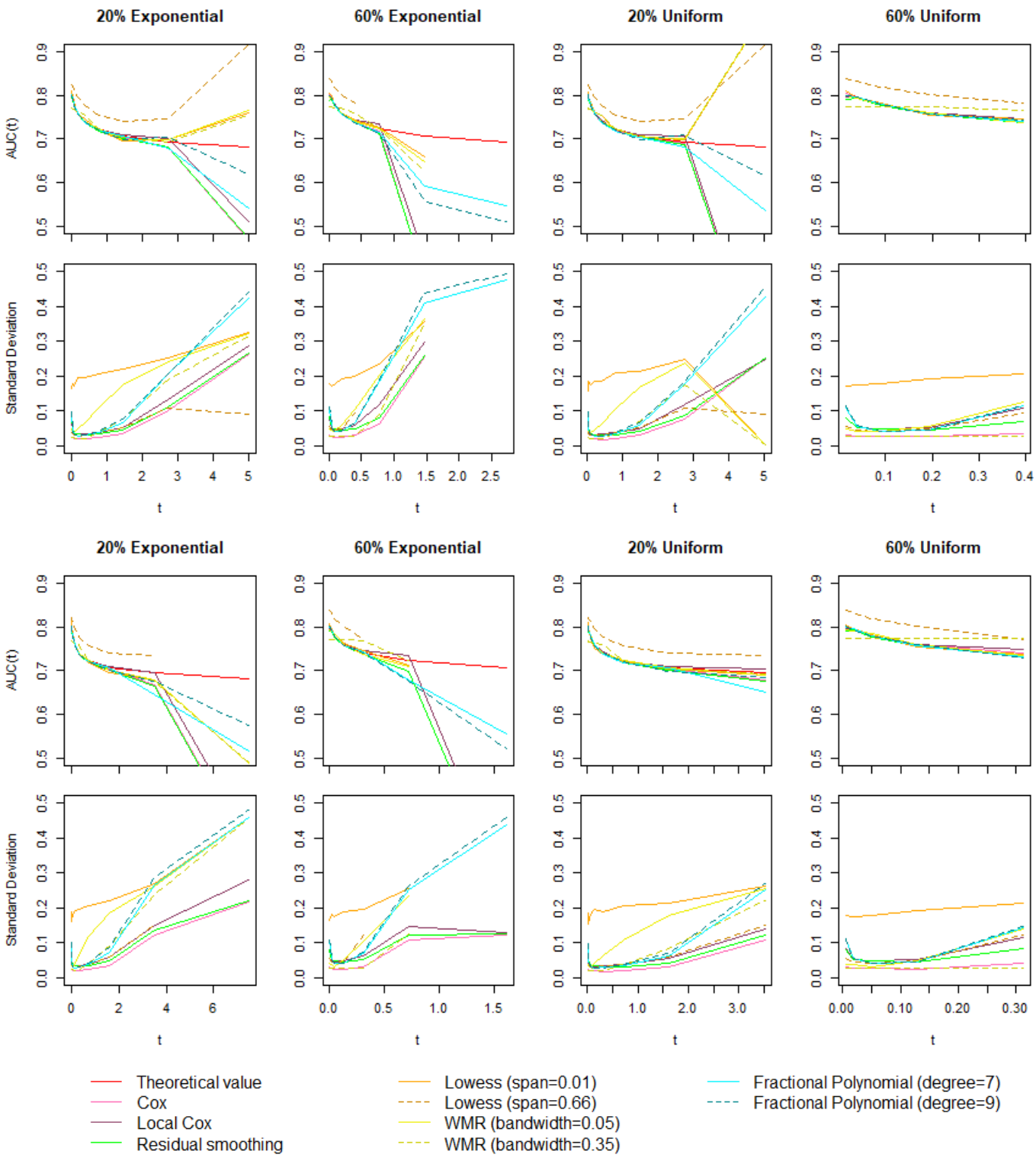


Figure B.2: Simulation scenario: Cox model based data with $\beta = 1.25$. Row 1, 2: increasing hazard. Row 3, 4: decreasing hazard.

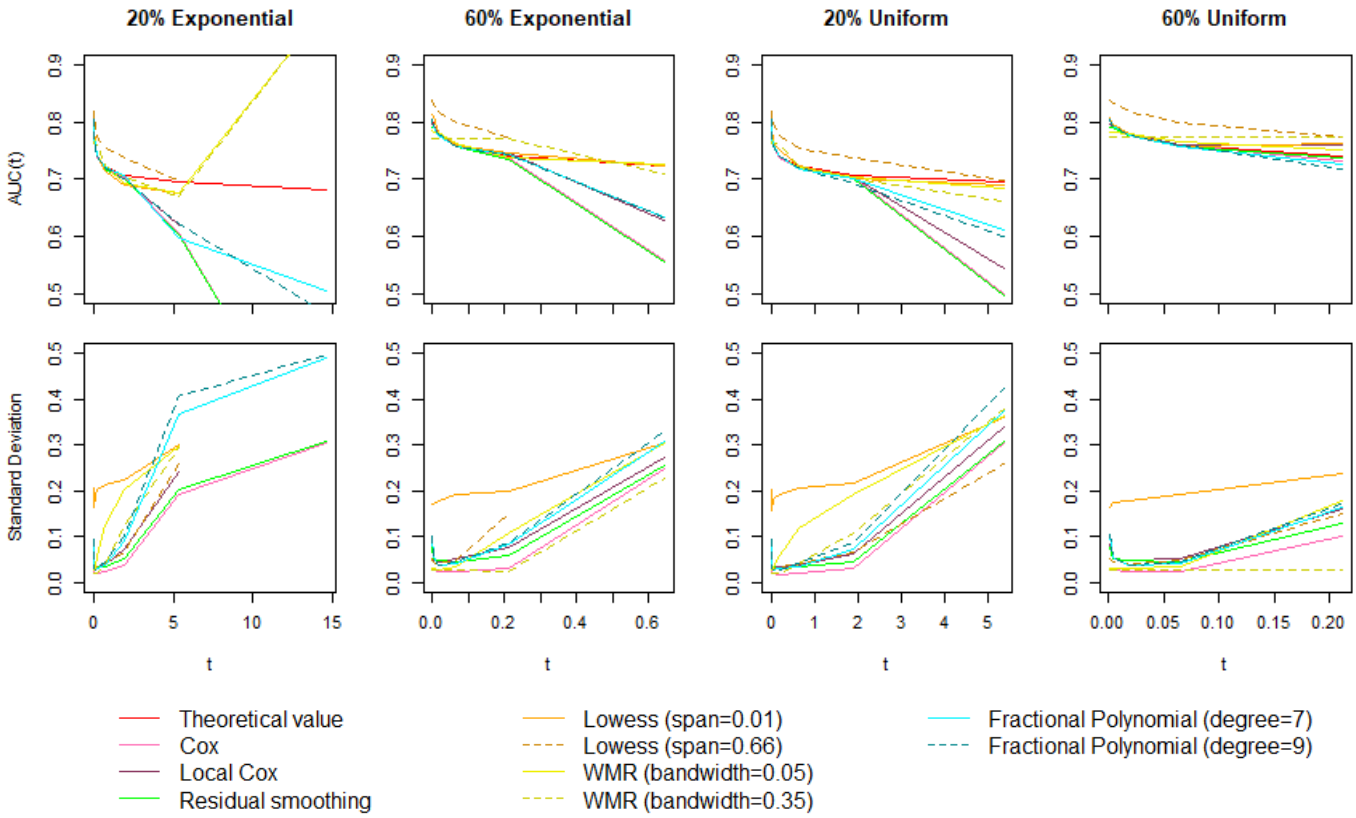


Figure B.3: Simulation scenario: Cox model based data with $\beta = 1.25$, decreasing hazard.

## B.2 $\beta = 0.4$

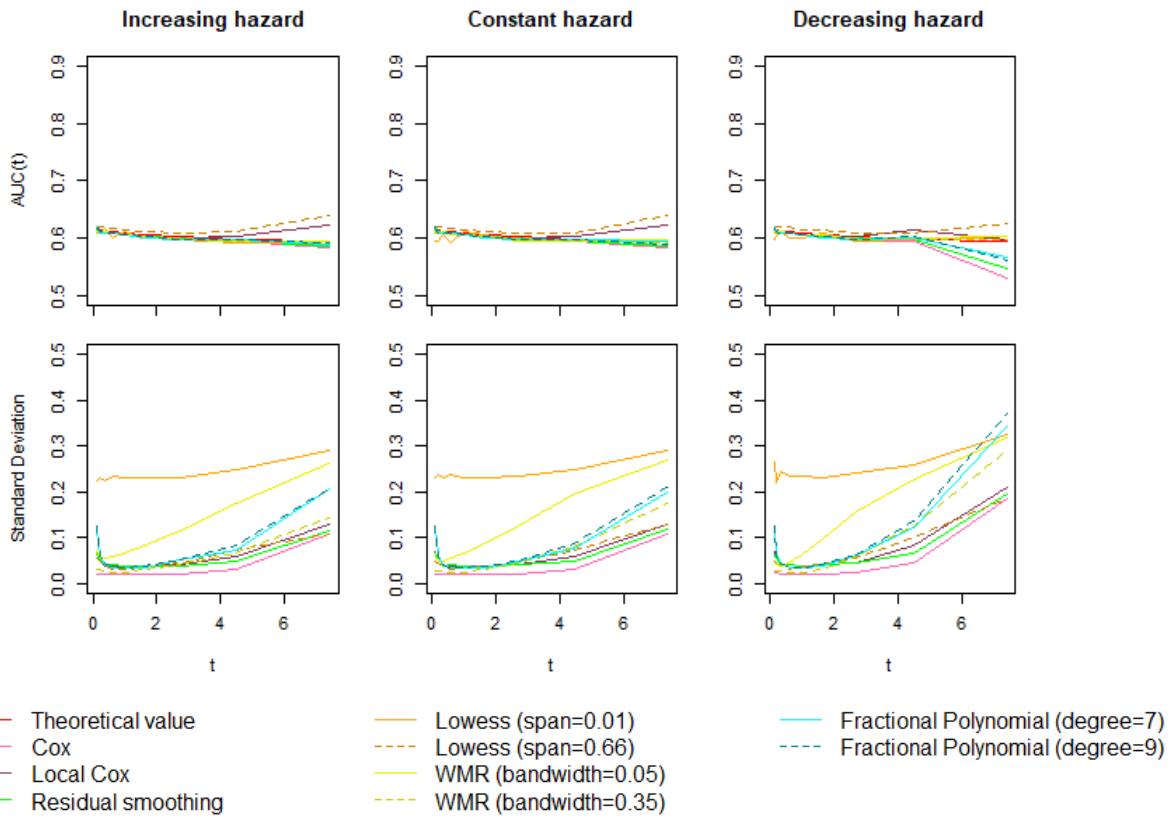


Figure B.4: Simulation scenario: Cox model based data with $\beta = 0.4$, 0% censoring.

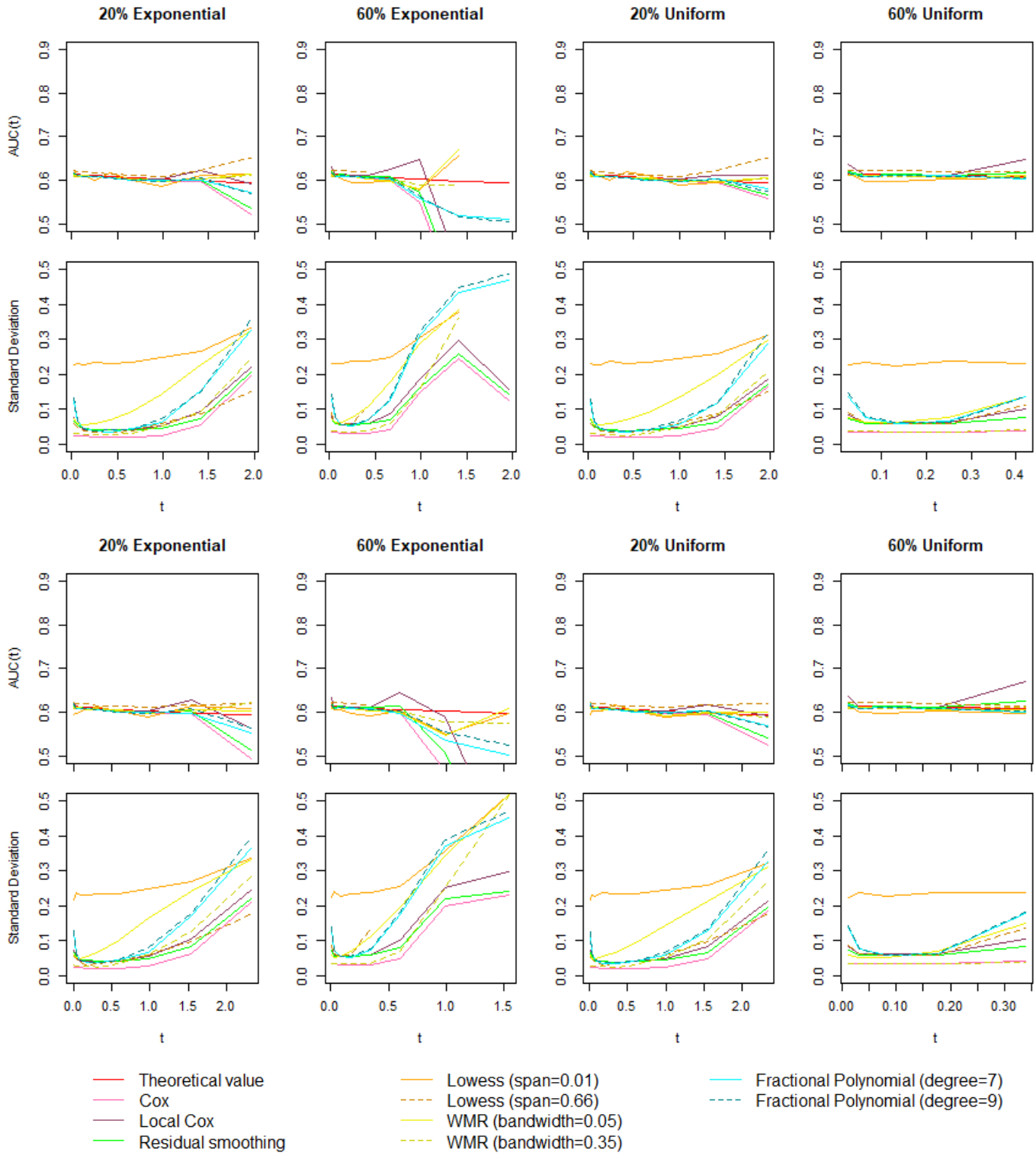


Figure B.5: Simulation scenario: Cox model based data with $\beta = 0.4$. Row 1, 2: increasing hazard. Row 3, 4: decreasing hazard.

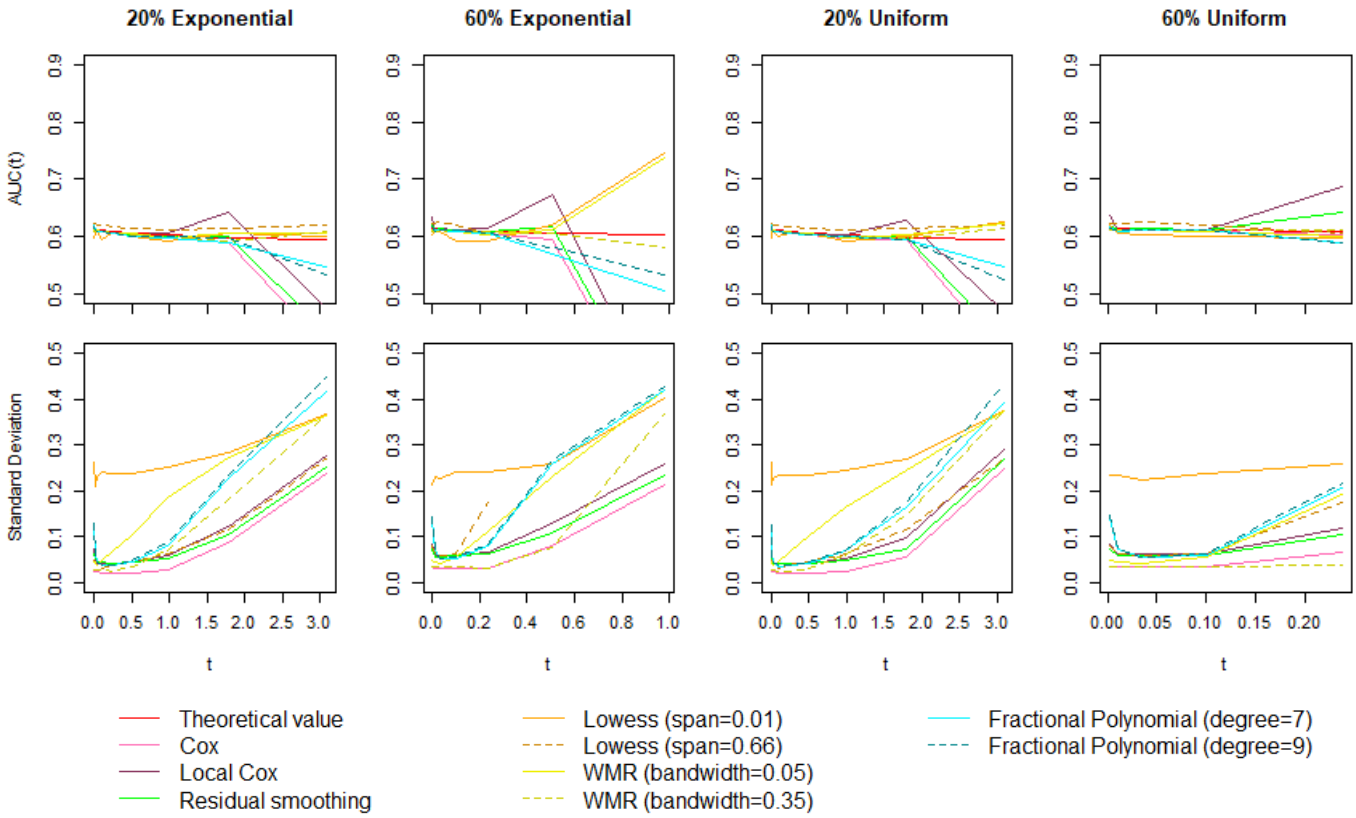


Figure B.6: Simulation scenario: Cox model based data with $\beta = 0.4$, decreasing hazard.