A comparison of methods for estimating Restricted Mean Survival Time

Yuqing Zhang (s1863444)

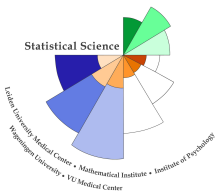
First supervisor: Prof. Dr. Hein Putter

External supervisor: Alberto Garcia Hernandez (Astellas)

MASTER THESIS

Defended on November 20^{th}, 2018

Specialization: Statistical Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

The Restricted Mean Survival Time (RMST) is a statistic that measures treatment effects which can be used as a replacement for the hazard ratio when the proportional hazards assumption is violated. The idea of RMST came from Irwin (1949) [5], and when combined with the formal definition of the survival function, RMST can be defined as the integral of survival function up to a time limit τ.

Several different methods for estimating the RMST are available. The Kaplan-Meier method and Cox PH model are the most commonly used methods in survival analysis, and they are also suitable for estimating RMST. This is done by first estimating the survival curve and then calculating the area under it to give an estimation of RMST. To allow a more general population of survival time distributions, a flexible parametric model was introduced by Royston and Parmar (2002) [4]. This flexible parametric model method followed the same method of estimating RMST as the Kaplan-Meier and Cox PH model: a survival function is estimated from the model, then a 15-point Gauss-Kronrod quadrature can be used to calculate the integral of the survival function, which allows estimation of RMST. The final option is a pseudo-observation method proposed by Anderson *et al.* (2004) [3]. This method first builds a pseudo-observation of RMST for each subject. Then, using the pseudo-observations of RMST as outcome variables, a generalized linear model can be built to describe the relationship between the covariates and RMST. A generalized estimating equation (GEE) method can then be used to estimate the parameters of the generalized linear model [8].

Comparisons between these methods under various simulation scenarios were conducted for this thesis. The Kaplan-Meier method is simple to calculate and performs well with early time limits and low censoring proportions. It is also faster to estimate RMST result than Cox model and flexible parametric model. However, this method lacks the ability to be adjusted for more covariates, so it is only suitable when estimating average RMST difference for a population. The unstratified Cox model performed well in datasets that satisfied the proportional hazards assumption. The stratified Cox model also performed well in our simulated non-proportional hazards datasets. The performance of the flexible parametric model method was similar to that of the Cox model, but it is more time-consuming in the integral calculation step. The pseudo-observation methods offered the shortest computation time among all four methods. However, when estimating RMST difference for a subject with given age and gender, the performance of the

pseudo-observation method was worse than either the Cox model or flexible parametric model.

Acknowledgements

I want to express my gratefulness to my thesis supervisors Prof. Dr. Hein Putter and Alberto Garcia Hernandez from Astellas. The teleconference with Alberto and bi-weekly meeting with Hein are really inspiring and informative. This thesis would not be conducted smoothly without advice and guidance from all supervisors. I want to thank Dr. Marta Fiocco for helping me with time arrangement and administration problems.

I wish to thank my friends and family for their accompany during the thesis.

Contents

1 Introduction **1**

1.1 Background 1

1.2 Aims and structure of this thesis 2

2 Methods **3**

2.1 Definition of Restricted Mean Survival Time (RMST) 3

2.2 Methods 4

2.2.1 Kaplan-Meier method (non-parametric) 5

2.2.2 Cox model (semi-parametric) 6

2.2.3 Flexible parametric model (parametric) 7

2.2.4 Pseudo-observation combined with linear model and Generalized estimating equation 10

2.3 Simulation scenarios 12

2.3.1 Simulation data setting 12

2.3.2 Model settings in the simulation study 15

2.3.3 Comparison scenarios 15

3 Results **17**

3.1 Non-proportional hazards situation 18

3.1.1 Influence of non-proportional hazards 18

3.1.2 Influence of censoring proportion 22

3.1.3 Influence of censoring distribution 23

3.1.4 Influence of baseline hazard shape 23

3.1.5 Estimation for a population with different ages and genders 24
3.1.6 Estimation for a subject with given ages and gender 24
3.2 proportional hazards situation 25
3.2.1 Influence of censoring proportion 26
3.2.2 Influence of censoring distribution 26
3.2.3 Influence of baseline hazard shape 27
3.2.4 Estimation for a population with different ages and genders 28
3.2.5 Estimation for a subject with given ages and genders 30
3.3 Computation time 30
3.4 Overall summary 32
3.4.1 Performance of methods under non-proportional hazards situations . 32
3.4.2 Performance of methods under proportional hazards situations . . . 33

4 Discussion **35**
4.1 Innovation and conclusion 35
4.2 Limitations and future work 36

Appendices **40**
A Appendix : code for thesis 40
A.1 brief instruction about how to use codes below 40
A.2 code for data generation 41
A.3 code for Kaplan-Meier method 46
A.4 code for Cox model method 48
A.5 code for Flexible parametric method 52
A.6 code for POGEE 55
B Appendix : Supplementary plots 58
B.1 standard deviation plots corresponding to RMST plots in Chapter 3 58
B.2 extra plots as supplement 62

Chapter 1

Introduction

1.1 Background

The hazard ratio is the most commonly used statistic to analyze the treatment effect in randomized clinical trials with time-to-event (survival) outcomes, and it is generally believed to be valid whenever the proportional hazards (PH) assumption is met. The proportional hazards assumption is quite strict; however, in actual analysis, this assumption is rarely checked. In recent years, however, researchers have found that the PH assumption may often be doubtful, particularly in randomized clinical trials with long follow-up in the field of oncology [**1**][**2**]. As the hazard ratio is believed to be biased and invalid under non-PH situations, to report treatment effect when the proportional hazards assumption is violated or doubtful, an alternative measurement, Restricted Mean Survival Time (RMST) may be considered. The RMST is expected to be a useful general measure to indicate the treatment effect under both PH and non-PH situations.

The definition of RMST is different from that of Mean Survival Time (MST) and is given with a specific time limit, τ. The RMST is restricted in time range $[0, \tau]$ to avoid the negative influences of the poorly determined right tail of a survival curve during estimation. The RMST is thus defined as the integral of survival function up to the chosen time limit τ.

Apart from the characteristic that it is free from PH assumptions, another advantage of RMST is its straightforward nature. An RMST difference result is equal to the RMST for the group taking treatment minus the RMST for the control group, assuming the result of the RMST difference is equal to δ, the time limit is τ, the event is death and time unit is years. An RMST difference result can thus be explained in this way: taking the treatment

will increase life expectancy by δ years over the next τ years as compared with not taking the treatment. This time domain explanation is generally easier to understand for patients and people unfamiliar with survival analysis.

1.2 Aims and structure of this thesis

This thesis will consider four estimation methods for RMST. The easiest way to estimate RMST is to calculate the area under a Kaplan-Meier survival curve. Kaplan-Meier is a non-parametric method used to estimate the survival function. Another method of estimating RMST is by integrating the survival function obtained using a Cox PH model, which is a semi-parametric method. In this thesis, Cox method is generalized to be stratified by a covariate in order to adjust for non-PH situations. Royston and Parmar (2002) [**4**] introduced a flexible parametric model where the RMST is estimated by integrating the survival function. This flexible parametric model is an extension of Weibull models, which use natural cubic splines to smooth the baseline cumulative hazard function. Andersen *et al.* (2004) [**3**] also suggested using pseudo-observations to estimate RMST for each subject. After generating pseudo-observations of RMST for every subject, a generalized linear model is used to model the relationship between the pseudo-observations of RMST and the covariates. The generalized estimating equation (GEE) method is then used to estimate the coefficients in the generalized linear model. For all four methods, ways to estimate the confidence interval are also briefly introduced. While additional methods have been developed for certain special kinds of survival data such as length-biased data [**11**], these methods are beyond the scope of this thesis.

A comparison between the Kaplan-Meier method, flexible parametric method, and pseudo-observation method was performed on real datasets by Royston and Parmar [**1**]. In this thesis, a comparison of all four methods of estimating RMST difference between arms in a simulation study is made in order to offer some suggestions for choosing appropriate methods for a given dataset.

The thesis is structured as follows: The Methods chapter contains the definition of RMST, a brief introduction of the four methods of estimating RMST, and the settings for the simulation scenarios. The Results section focuses on the simulation study results, while the Discussion chapter highlights key results, discusses the limitations of the study and offers suggestions for possible further work.

Chapter 2

Methods

2.1 Definition of Restricted Mean Survival Time (RMST)

Mean survival time (MST) measures expected life time and can be used to quantify treatment effect in a time domain which can be easily interpreted as life expectancy when the event is death. For survival data, if the proportion of censoring is high, the estimation of MST will not be satisfactory. Due to right-censoring, the right tail of the survival curve and accordingly the MST may be ill-determined. In such cases, the RMST is introduced to estimate the mean life limited to reaching time point τ, which defined as μ_τ. Here, μ_τ is the mean of $min[T_i, \tau]$. The time point τ, is defined by the researcher according to the research goals. Normally, the choice of τ is connected to the censoring and survival distribution, as the purpose of setting a time limit τ is to avoid the ill-determined problem at later time area to some extent.

The idea of RMST was developed in demography by Irwin in 1949 [**5**]. An expectation of life limited to time τ is defined as

$$\frac{1}{l_0}\int_0^\tau l_x dx,$$

where l_0 is the fixed base number presumed to be born at any moment in time and $l_x = l_0 p_0 p_1 ... p_{x-1}$. Each p_i is the probability of surviving for one time unit at time i.

The definition given above is a little bit different from the survival definition currently used. The survival function, $S(t)$, present the probability for an individual to survive at time t. Thus, $S(r) = p_0 p_1 p_2 ... p_{r-1} = \frac{l_r}{l_0}$. The original definition of RMST by Irwin (1949)

[**5**] can thus be transformed into

$$\frac{1}{l_0}\int_0^\tau l_x dx = \int_0^\tau \frac{l_x}{l_0} dx = \int_0^\tau S(x)dx$$

.

This thesis uses

$$\mu_\tau = E[min(T,\tau)] = \int_0^\tau S(t)dt$$

as the definition of Restricted Mean Survival Time, where τ is the predetermined time limit and $S(t) = P(T > t)$ is the survival function. The RMST can thus be described as the area under the survival curve in time range $[0, \tau]$ (Figure 2.1).

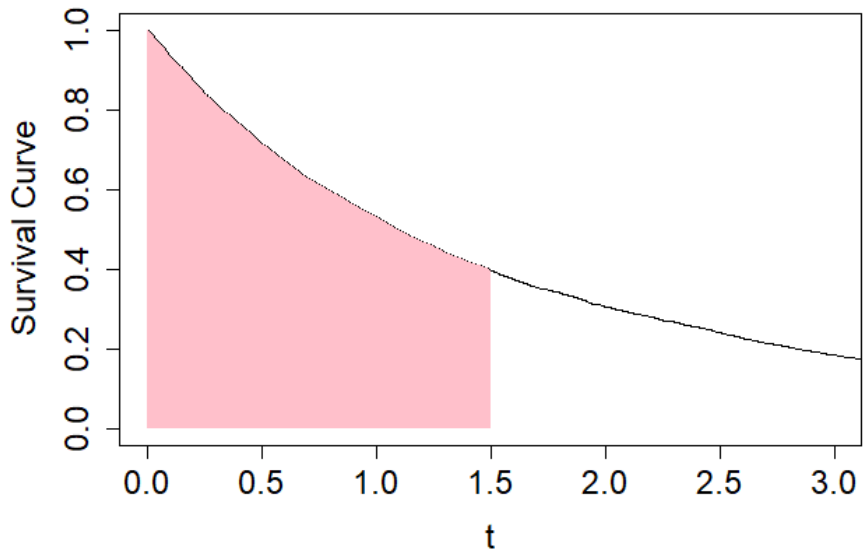


Figure 2.1: RMST as the area (pink shading) under the survival curve when τ is set equal to 1.5.

2.2 Methods

This section introduces four different methods for estimating RMST and its confidence interval. The first three methods discuss non-parametric, semi-parametric, and parametric methods for estimating the survival function. Based on the survival function estimation results, the integral of the survival function up to chosen time limit, τ, will be treated as the estimation of RMST. The final method, pseudo-observation combined with a linear model and GEE, offers another approach to estimating RMST by estimating a pseudo-

value of the RMST for each subject first and then building a regression model that has the estimated RMST as an outcome variable.

2.2.1 Kaplan-Meier method (non-parametric)

The Kaplan-Meier estimator of the survival function is defined in [**10**] as:

$$\widehat{S}(t) = \begin{cases} 1 & t < t_1 \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}] & t_1 \leq t, \end{cases}$$

where d_i is the number of events at time point i, Y_i is the number at risk at time point i and t_i is the time point with an event or censoring.

The survival curve of a Kaplan-Meier estimated survival function is a step function, which jumps at the time points where an event occurs. The RMST result, which is represented by the area under the survival curve can be calculated accurately as

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt = \sum_{t_i \leq \tau}(t_i - t_{i-1})S_{i-1},$$

where t_0 equal to 0 and S_{i-1} is the survival probability between t_{i-1} and t_i.

The $100(1-\alpha)\%$ confidence interval for RMST under Kaplan-Meier method is equal to [**10**]:

$$\hat{\mu}_\tau \pm Z_{1-\alpha/2}\sqrt{\hat{V}[\hat{\mu}_\tau]},$$

where the variance of the survival function is estimated using Greenwood's formula [**10**]:

$$\hat{V}[\hat{\mu}_\tau] = \sum_{i=1}^{D}[\int_{t_i}^{\tau} \hat{S}(t)dt]^2 \frac{d_i}{Y_i(Y_i - d_i)},$$

where D is the number of event time points that occur prior to the time limit τ.

This method is a non-parametric method which is easy to estimate the survival function and estimate RMST. However, the estimation of RMST under this method is based on average survival which makes it impossible to adjust for additional covariates.

The $survfit$ function in the *survival* package for R can be used to estimate survival functions using Kaplan-Meier method.

2.2.2 Cox model (semi-parametric)

The Cox PH model is as

$$h(t \mid Z) = h_0(t) \exp(\sum_{k=1}^{P} \beta_k Z_k),$$

where h_0 is the baseline hazard. Under the Cox proportional hazards model, it is possible to estimate the survival function of a given covariates set Z_0. Let b be the partial maximum likelihood estimators of coefficients β. Let $t_1 < t_2 < ... < t_D$ indicate the time points where events occurred, it is possible to define

$$W(t_i; b) = \sum_{j \in R(t_i)} \exp(\sum_{k=1}^{p} b_k Z_{jk}),$$

where $R(t_i)$ is the set of individuals at risk at time t_i and b_h is the partial maximun likelihood estimator of β_h, p is the total number of covariates and Z_{jh} indicate the value of covariate h for an individual j.

Let d_i be the number of event at time t_i. Breslow's estimator of the cumulative baseline hazard rate (in [**10**]) is as follows

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i; b)},$$

which can be computed using the *basehaz* function in *R*. As an alternative to Breslow's method, the estimator proposed by Kalbfleisch and Prentice (1973) [**6**] can be used. When there are no tied events at each time point, this estimator is given by

$$\tilde{H}_0(t) = \sum_{t_i \leq t} [1 - (1 - \frac{\delta_i exp(b^t Z_i)}{W(t_i; b)})^{exp(-b^t Z_i)}].$$

The simulation study in this thesis use Breslow's method.

The estimation of the survival function for the baseline group is given by

$$\hat{S}_0(t) = \exp[-\hat{H}_0(t)].$$

For individuals with covariates $Z = Z_0$, the estimator of the survival function will be

$$\hat{S}(t \mid Z = Z_0) = \hat{S}_0(t)^{\exp(b^t Z_0)}.$$

The estimated survival function again operates as a step function, just as in the Kaplan-Meier results. The RMST is also calculated as the area below the survival curve in the

same way as in the Kaplan-Meier method using

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt = \sum_{t_i \leq \tau} (t_i - t_{i-1}) S_{i-1}.$$

To build the confidence interval for the RMST, we use the bootstrap method. The bootstrap method replicates the RMST estimation procedure thousands of times based on different sample data selected from the whole dataset. The standard deviation of the bootstrap RMST estimations will be treated as the estimation of the standard error of RMST. Then, the confidence interval can be estimated. There are couple of methods estimating bootstrap confidence interval. Carpenter and Bithell (2000) conducted a comparison between these methods [**7**]. And a suitable bootstrap method can thus be chosen based on computation complexity and applicability for different simulation scenarios. Here, in our later simulation study, we use the basic (non-studentized) bootstrap method which can be applied in *R* using function *boot.ci* from the *boot* package.

To generalize to non-PH situations, there are several possible methods. One simple way is to use a Cox model stratified by the covariate that violates the PH assumption. This generalized model was used in the simulation studies.

The *survfit(coxph)* function in the *survival* package in *R* can compute the estimated survival function based on the Cox model.

2.2.3 Flexible parametric model (parametric)

Another method to estimate RMST is to follow a parametric approach that essentially fitting PH or non-PH models similar to the ones proposed in section 2.2.2. Different from Cox model, a parametric approach will not leave the baseline hazard function unspecified.

The basic parametric models use the Weibull or exponential distribution to fit the baseline hazard. However, real survival data often do not follow such a simple pattern. A more flexible model is needed.

Royston and Parmar (2002) [**4**] proposed a flexible parametric model as an extension of Weibull models using natural cubic splines to smooth the baseline cumulative hazard function. The model allows more distributions other than Weibull and can be extended for a non-PH scenario.

The basic model for these flexible parametric models is based on the PH assumption. And this kind of model can be described as a transformation of the survival function with

a link function $g(\cdot)$:

$$g[S(t;Z)] = g[S_0(t)] + \beta^T Z,$$

where $S_0(t)$ is the baseline survival function. Royston and Parmar (2002) [**4**] use natural cubic splines to model $g[S_0(t)]$ with the link function family $g(\cdot)$ suggested by Aranda-Ordaz (1981):

$$g(x;\theta) = \ln \frac{x^{-\theta} - 1}{\theta},$$

where $\theta \to 0$ correspond to the proportional hazards model.

This method start with Weibull models. Suppose T is a random variable with a Weibull distribution, scale parameter λ, and shape parameter α; let $x = \ln(t)$. This gives

$$\ln H(t) = \ln[\lambda t^{\alpha}] = \alpha x + \ln(\lambda) = \gamma_0 + \gamma_1 x,$$

which is linear in x. If the distribution of T changes from Weibull, then $\ln H(t)$ will be related to x by means of a non-linear function $s \equiv s(x;\gamma)$.

We estimate the survival function by smoothing the cumulative baseline hazard function. For the basic spline model with fixed covariate vector Z under the PH assumption, we have

$$g[S(t;Z)] = \ln[-lnS(t;Z)] = \ln H(t;Z) = \ln H_0(t) + \beta^T Z = s(x;\gamma) + \beta^T Z.$$

We use the natural cubic spline to smooth the $s(x;\gamma)$. The natural cubic splines are linear beyond boundary knots k_{min} and k_{max}, while between boundary knots, m internal knots are specified such that $k_{min} < k_1 < ... < k_m < k_{max}$. A natural cubic spline for $s(x;\gamma)$ may thus be written as:

$$s(x;\gamma) = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + ... + \gamma_{m+1} v_m(x),$$

the j-th basic function is defined for $j = 1, 2, ..., m$ as

$$v_j(x) = (x - k_j)_+^3 - \lambda_j (x - k_{min})_+^3 - (1 - \lambda_j)(x - k_{max})_+^3,$$

where $\lambda_j = (k_{max} - k_j)/(k_{max} - k_{min})$ and $(x - a)_+ = max(0, x - a)$.

The complexity of the smoothed curve depends on the degrees of freedom which are equal to $m + 1$. When m equal to 0 ($df = 1$), no knots will be specified, and the baseline distribution is thus Weibull. When choosing m, Royston and Parmar (2002) [**4**] suggested observing the AIC value for models with different number of m and choosing the model

with the smallest AIC value.

The above models based on PH assumption can be extended to include non-proportional scaling for some subset of the covariates. For Cox PH model, one way to deal with non-proportionality is making regression coefficients depend on a predefined function $f(t)$ of time. The formulation of $f(t)$ can be a simple fixed function or more flexible spline functions and step functions. Based on this approach, Royston and Parmar (2002) [**4**] use a closely related approach to extend the basic flexible parametric model.

A PH spline model with a single covariate z_1 and a single knot can be written as

$$\ln H(t; z_1) = \gamma_0 + \gamma_{10}x + \gamma_{20}v_1(x) + \beta_1 z_1.$$

A extension of the PH spline model to allow a time-dependent log cumulative hazard ratio for z_1 is as

$$\ln H(t; z_1) = \gamma_0 + (\gamma_{10} + \gamma_{11}z_1)x + (\gamma_{20} + \gamma_{21}z_1)v_1(x) + \beta_1 z_1.$$

When generalising the above model for m internal knots and covariates set $\mathbf{z}$, the spline model is:

$$g[S(t; Z)] = \ln H(t; Z) = \gamma_0 + \gamma^T \mathbf{v}(x) + \beta^T \mathbf{z},$$

where $\mathbf{v}(x) = (x, v_1(x), ..., v_m(x))^T$. Assuming that the first k covariates in covariate set $\mathbf{z}$ are under non-proportional scaling, then the jth element of γ is $\gamma_j = \gamma_{j0} + \sum_{l=1}^{k} \gamma_{jl} z_l$.

After constructing the model, full maximum likelihood can then be used to estimate γ and β. According to

$$g[S(t; Z)] = \ln[-\ln S(t; Z)] = \ln H(t; Z) = \ln H_0(t) + \beta^T Z = s(x; \gamma) + \beta^T Z,$$

the estimated survival function can be produced by using the estimated γ and β. The estimated RMST will thus be the integral of $S(t; Z)$.

The *flexsurv* package in *R* can be used to construct the model and obtain the survival result for a given covariate's value and time point. To estimate the integration of the survival curve, the 15-point Gauss-Kronrod quadrature is used; this method can be implemented in *R* by utilizing the *integrate* function [**9**].

A bootstrap method can also be used here to estimate the standard error and the confidence interval of RMST estimation as described in section 2.2.2; however, when using *R*, the time taken by the bootstrap method under this parametric model is very long. The delta method may be another choice. The delta method may improve the computing

time, but enjoy more complex codes. In R, conducting delta method require typing in the targeted statistic as a function of parameter estimates. In our cases, the exact function for RMST is a complicated one combined by a natural cubic spline model and a 15-point Gauss-Kronrod quadrature estimation. The workload of coding for delta methods is much heavier compared to bootstrap.

2.2.4 Pseudo-observation combined with linear model and Generalized estimating equation

Andersen *et al.* (2004) [**3**] used pseudo-observations to estimate RMST. Unlike the previous three methods, which all estimate the RMST by calculating the area under the estimated survival curve, the pseudo-observation method calculates a pseudo-value for the RMST directly for every subject. Combined with a generalised linear model and generalized estimating equation (GEE), this allows estimation of RMST for a population based on the given values of covariates.

The pseudo-observation is defined by the following steps:

(1) Set $X_i, i = 1, 2, ..., n$ as independent and identically distributed random variables.

(2) θ is a parameter of interest following the form $\theta = E(f(X_i)) = \int f(x)P(dx)$. Assume an unbiased estimator $\hat{\theta}$ for θ.

(3) Define the conditional expectation as $\theta_i(Z_i) = E[f(X_i) \mid Z_i]$, where $Z_i, i = 1, 2, ..., n$ are independent and identically distributed covariates.

(4) The i-th pseudo-observation is

$$\hat{\theta}_i = n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}^{-i}.$$

Here, $\hat{\theta}^{-i}$ is the 'leave one out' estimator for θ based on $X_j, j \neq i$.

The most distinctive characteristic of this pseudo-observation method is that the parameter of interest, θ, is constructed for every individual. When constructing pseudo-observations for RMST, the parameter of interest θ is equal to μ_τ. The function f is given by $f(X) = min[X, \tau]$ for $\tau > 0$.

The i-th pseudo-observation of RMST is therefore

$$\hat{\mu}_{\tau i} = n \cdot \hat{\mu}_\tau - (n-1) \cdot \hat{\mu}_\tau^{-i} = n \cdot \int_0^\tau \hat{S}(t)dt - (n-1) \cdot \int_0^\tau \hat{S}^{-i}(t)dt = \int_0^\tau [n\hat{S}(t) - (n-1)\hat{S}^{-i}(t)]dt,$$

where $i = 1, 2, ..., n$; $\hat{S}(t)$ is survival function estimated by Kaplan-Meier method; and $\hat{S}^{-i}(t)$ is the 'leave-one-out' Kaplan-Meier estimator of survival function based on observations $X_j, j \neq i$.

Andersen and Pohar Perme (2009) [**8**] indicated that the pseudo-observation of RMST for an individual i is equal to the event time X_i when the data is without censoring. In the presence of censoring, the pseudo-observations of RMST for a censored individual will always be larger than X_i and significantly larger than the pseudo-observations of RMST for uncensored individuals with similar observed times.

Once pseudo-observations of RMST for each individual are obtained, a regression model for the RMSTs can be fitted. Considering generalised linear models, the function becomes

$$g(E[f(X) \mid Z]) = \beta_0 + \sum \beta_j Z_j,$$

where g is a given link function.

A generalized estimating equation (GEE) is used to estimate the parameters [**8**]. The estimating equation is as

$$U(\beta) = \sum_i U_i(\beta) = \sum_i (\frac{\partial}{\partial \beta} g^{-1}(\beta^T Z_i^*))^T V_i^{-1} (\hat{\mu}_{\tau i} - g^{-1}(\beta^T Z_i^*)),$$

where Z_i^* include indicator of time point and covariates Z_i and V_i is the covariance matrix. The parameters β are estimated by solving $U(\beta) = 0$. For a given set of covariates, a prediction of RMST can thus be obtained according to the model fitting results.

A bootstrap method is used under this method to estimate the standard error of RMST to build confidence intervals. The choice of the bootstrap method is the same as section 2.2.2 and 2.2.3.

The advantage of this method is based on the fact that every individual is given a pseudo-observation for the RMST with the defined time limit. When fitting a regression model treating pseudo-observations of RMST as a response variable, the estimated coefficients will give information directly about how RMST depends on covariates.

The *pseudo* package and the *geepack* package can be used to compute pseudo-observation results of RMST and build the regression model in *R*.

2.3 Simulation scenarios

In this section, the simulation scenarios used for the comparison are described. The methods used to build the proportional hazards and non-proportional hazards datasets will be introduced; then, the model fittings used in the simulation study are described. Finally, the comparison scenarios will be presented within a table showing the combinations of different data types, censoring settings and hazard shapes.

2.3.1 Simulation data setting

Simulation data was constructed for two treatment groups with other two covariates, age and gender. Covariate age is a continous variable uniformly distributed in the domain $[20, 60]$ and covariate gender is a binomial variable where a value of 0 indicates male and 1 indicates female.

The baseline hazard function is assumed to follow the Weibull distribution where the hazard function of the Weibull distribution is $h(t) = \alpha\lambda t^{\alpha-1}$, α is the shape parameter, and λ is the scale parameter. By changing α and λ, different shapes of hazard function can thus be produced (Figure 2.2). Here, I use $\alpha = 1$ and $\lambda = 1$ to form a constant hazard, $\alpha = 1.25$ and $\lambda = 0.9$ indicate increasing hazard, and $\alpha = 0.75$, $\lambda = 1.15$ indicate decreasing hazard. When using these three set of parameter settings, the $mean(E(X))$ of Weibull distribution will approximately be the same, namely $1(\alpha = 1, \lambda = 1)$.

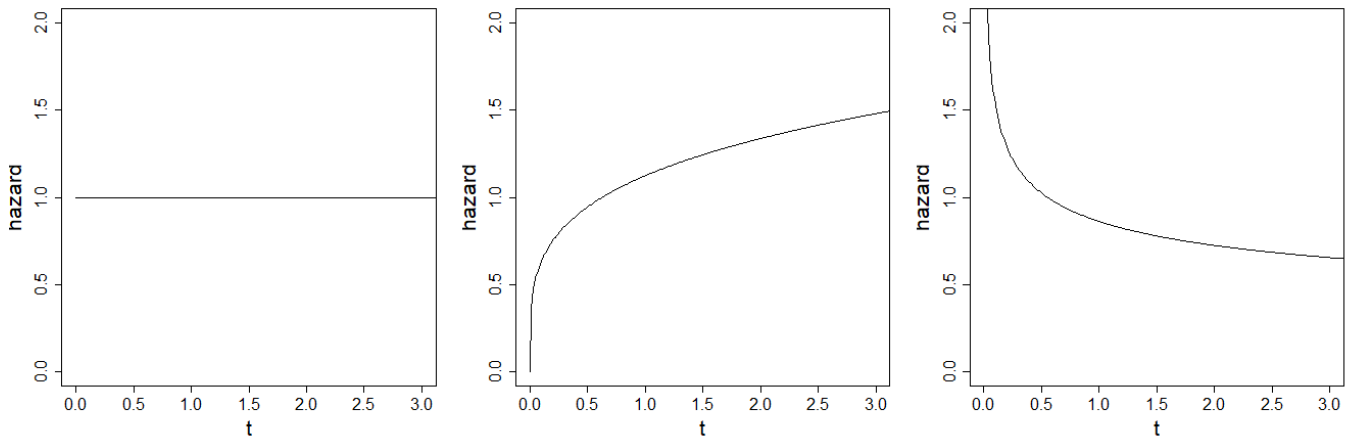


Figure 2.2: Different shape of Weibull baseline hazard in time range $[0, 3]$. Left panel: $\alpha = 1$, $\lambda = 1$; Middle panel: $\alpha = 1.25$, $\lambda = 0.9$; Right panel: $\alpha = 0.75$, $\lambda = 1.15$

The cumulative hazard invertion method (Bender *et al.* (2005) [**12**]) is used to generate the event time for each individual. A sample u is taken from uniform distribution $U(0, 1)$

and treated as the survival probability for individual i. The event time for individual i is $T_i = S_i^{-1}(u)$ where the $S_i^{-1}(u)$ is the inverse survival function. Based on $S_i(u) = exp(-H_0(u)exp(Z_i^T\beta))$ and $H_0(u) = \lambda u^\alpha$, we have $T_i = S_i^{-1}(u) = [\frac{-log(u)}{\lambda exp(Z_i^T\beta)}]^{1/\alpha}$.

Simulation data was constructed both under the proportional hazards assumption and for a non-proportional hazards situation.

For the proportional hazards situation, the control group was treated as the baseline group following a Weibull distribution with the log hazard ratio between treatment group and control group being equal to -0.5. The log hazard ratio for covariates age and gender were also set as -0.5. The Kaplan-Meier survival curves for the three PH situations used in the simulation study are showed in Figure 2.3.

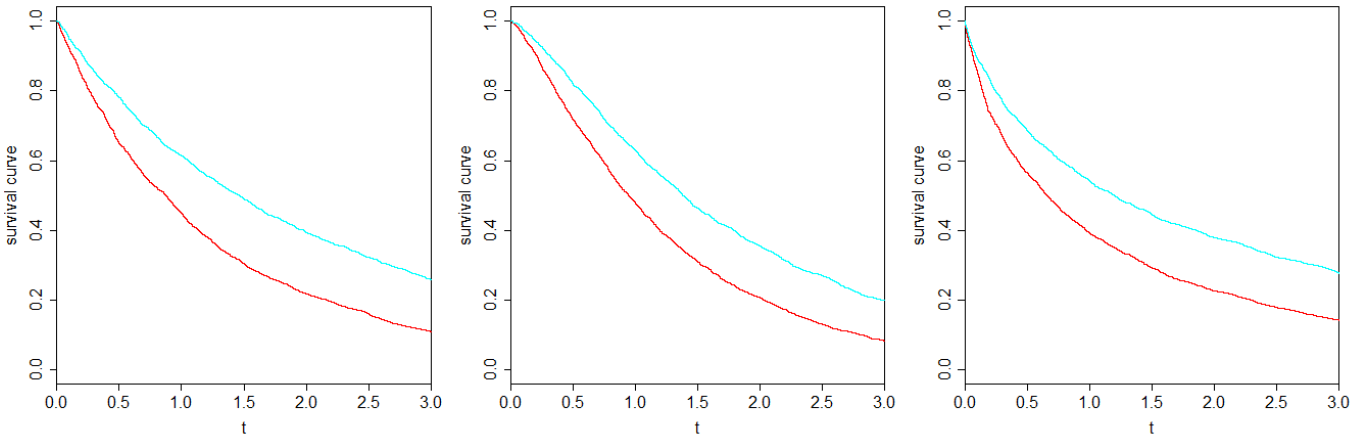


Figure 2.3: Survival curves for two treatment groups that meet the proportional hazards assumption. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard. Blue line: treatment group; Red line: control group.

When different baseline hazard shapes are assigned to the two treatment groups, the proportional hazards assumption between treatment groups is violated. A total of three combinations of baseline hazard shapes was considered for the non-PH scenarios: constant and increasing, constant and decreasing and decreasing and increasing. These survival curves are showed in Figure 2.4.

In R, the *simsurv* package can be used to apply this cumulative hazard invertion method.

After generating covariates Z_i and an event times T_i for every individual i, the censoring time C_i was respectively generated following an exponential distribution and uniform distribution. The final observation time t_i for individual i is $\min[T_i, C_i]$, and any individual

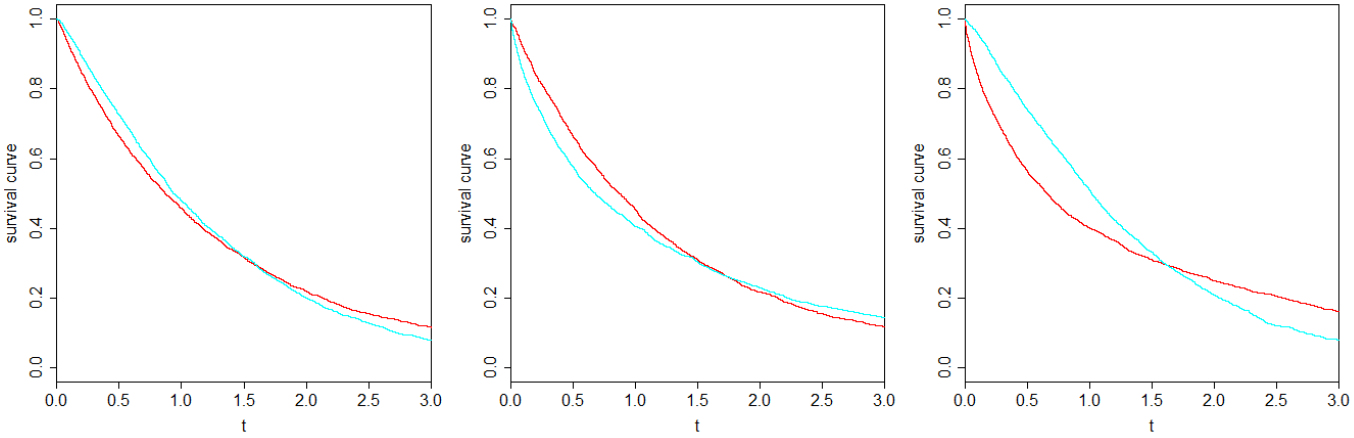


Figure 2.4: Survival curves for two treatment groups that violate the proportional hazards assumption. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group. Blue line: treatment group; Red line: control group.

with $t = C_i$ is censored. The trial end was set at three years, and thus subjects with an event or censoring at time points greater than 3 were right censored at 3. The exponential distribution was proposed, with rate parameters set to ensure that, overall $(25 \pm 0.5)\%$ and $(50 \pm 0.5)\%$ of subjects in the data set were censored. A uniform distribution was also proposed, with parameters set to ensure $(25 \pm 0.5)\%$ of the subjects were censored. Under our simulation settings, a $(50 \pm 0.5)\%$ uniform censoring is hard to reach. When $C_i \sim U[a, b]$, there are two ways to increase the proportion of censoring. One option is to decrease the number of a towards 0 and another option is to decrease b towards trial end time. Under our simulated data, making $C_i \sim U[0, 3]$ will gain the highest proportion of censoring, but it is still smaller than $(50 \pm 0.5)\%$. To keep consistency with exponential censoring, we only include the $(25 \pm 0.5)\%$ uniform censoring results. The parameter choices for censoring distribution are shown in Table 2.1.

Table 2.1: Choice of rate parameters and uniform parameters for different simulation scenarios

	25% exp censoring	50% exp censoring	25% uniform censoring
PH, constant hazard	rate = 0.08	rate = 0.58	U(1, 5)
PH, increasing hazard	rate = 0.125	rate = 0.55	U(1, 4)
PH, decreasing hazard	rate = 0.045	rate = 0.65	U(1, 8)
non-PH, constant; increasing	rate = 0.2	rate = 0.725	U(0.5, 4)
non-PH, constant; decreasing	rate = 0.18	rate = 0.85	U(0.5, 4)
non-PH, decreasing; increasing	rate = 0.2	rate = 0.8	U(0.5, 4)

2.3.2 Model settings in the simulation study

Kaplan-Meier method is non-parametric, it can not be adjusted for more covariates other than treatment. We use Kaplan-Meier model to estimate RMST result for each treatment group separately and this setting will be used under all comparison scenarios when the Kaplan-Meier method is applicable. For Cox model, four different settings are used: a Cox PH model with only treatment as covariate; a Cox PH model adjusted for covariates treatment, age and gender; a Cox model stratified by treatment and not adjusted for other covariates; a Cox model stratified by treatment and adjusted for covariates age and gender. For flexible parametric model, four settings are used similar to the Cox model: a flexible parametric model under PH assumption and only include covariate treatment; a flexible parametric model under PH assumption and adjusted for covariates treatment, age and gender; a flexible parametric model allow non-proportionality for treatment and not include other covariates; a flexible parametric model allow non-proportionality for treatment and include covariates age and gender. We use 3 internal knots when applying the natural cubic spline. The pseudo-observation combined with GEE method is free from PH assumption. Two settings are used in the simulation study. One only includes treatment in the regression model, and the other include treatment, age, and gender in the regression model.

2.3.3 Comparison scenarios

Comparison scenarios are combinations of censoring type, hazard shape, and data assumption (see Table 2.2). For each scenario, 1000 data sets with 400 subjects, 200 subjects for each treatment group, were generated. The RMST was estimated under six time limits, $\tau = \{0.5, 1, 1.5, 2, 2.5, 3\}$, to investigate performance of methods at early time limits and later time ranges where survival curve may be ill-determined.

Table 2.2: Comparison scenarios.

	no censoring	25% exp censoring	50% exp censoring	25% unif censoring
difference of RMST between	cons	cons	cons	cons
treatment groups, PH	incr	incr	incr	incr
(one baseline hazard)	decr	decr	decr	decr
difference of RMST between	cons; incr	cons; incr	cons; incr	cons; incr
treatment groups, non-PH (two baseline	cons; decr	cons; decr	cons; decr	cons; decr
hazard shapes for two treatment groups)	decr; incr	decr; incr	decr; incr	decr; incr

'cons' stands for constant baseline hazard, 'incr' stands for increasing baseline hazard, 'decr' stands for decreasing baseline hazard. For non-PH situations, the first baseline hazard information is for the control group and the second baseline hazard information is for the treatment group.

The methods were compared estimating the difference in RMST between the two treatment groups. For each comparison scenario, the final estimated result for the difference

of RMST was the mean of $1,000$ estimations; the standard deviation was also computed over $1,000$ estimations. A theorectical true value is included in all comparison scenarios as standard.

For every scenario, two different set of covariates were used for comparison. When only covariate treatment is included in the models, the RMST difference estimation was the average RMST difference for a population with different ages and gender. When including all three covariates treatment, age and gender in the models, we can estimate the difference of RMST for a subject with given age and gender. In our simulation study, we set the age and gender as 50 years old and male.

Chapter 3

Results

This chapter presents the simulation results for the different scenarios described in Chapter 2. The various methods will be compared from different angles within proportional hazards data sets and non-proportional hazards data sets. A simple comparison of computation time for all methods will also be conducted.

All the RMST difference estimation results in this chapter will be the RMST estimations for the treatment group minus the RMST estimation for the control group. The estimation results in all plots are mean results over 1,000 simulated datasets. The standard deviations over the 1,000 simulated datasets will also be computed.

The pseudo-observation combined with a linear model and GEE method will be written as POGEE in the legends of plots and further analysis, FP in the plot legend stands for flexible parametric model and KM stands for Kaplan-Meier method.

In all plots, the model settings are shown in the legend. For POGEE, Cox, and FP, the term 'simple' means that only the treatment is included in the models; while 'adjusted' means the methods includes treatment, age, and gender in the model. For Cox model and FP model, the term 'PH' means the model is based on the PH assumption, while 'nPH' means the model is extended for nonproportionality. The Cox model is stratified for treatment enable to allow nonproportionality and FP model is adding time-dependent functions for treatment effect to allow nonproportionality.

Different color in plots present the theoretically true values and estimation results of different methods. The color for each method is indicated as the legend in each plot.

3.1 Non-proportional hazards situation

As the RMST statistic acts as a replacement of the hazard ratio mainly where the proportional hazards assumption is violated, the comparison results under non-proportional hazards scenarios are the first to be demonstrated.

3.1.1 Influence of non-proportional hazards

This subsection focuses on the effects of non-proportional hazards on RMST difference results. And examines what happens if the verification of the proportional hazards assumption is neglected.

The RMST difference between groups can be illustrated graphically as the area between the two survival curves for the treatment group and control group. As shown in Figure 2.3, when datasets meet the PH assumption, the curves do not cross, and the curve for the treatment group (blue line) is always above that for the control group (red line). It is thus obvious that the area between the curves increases monotonically with an increase in the time limit. Figure 3.1 presents a clear view of this monotonic trend.

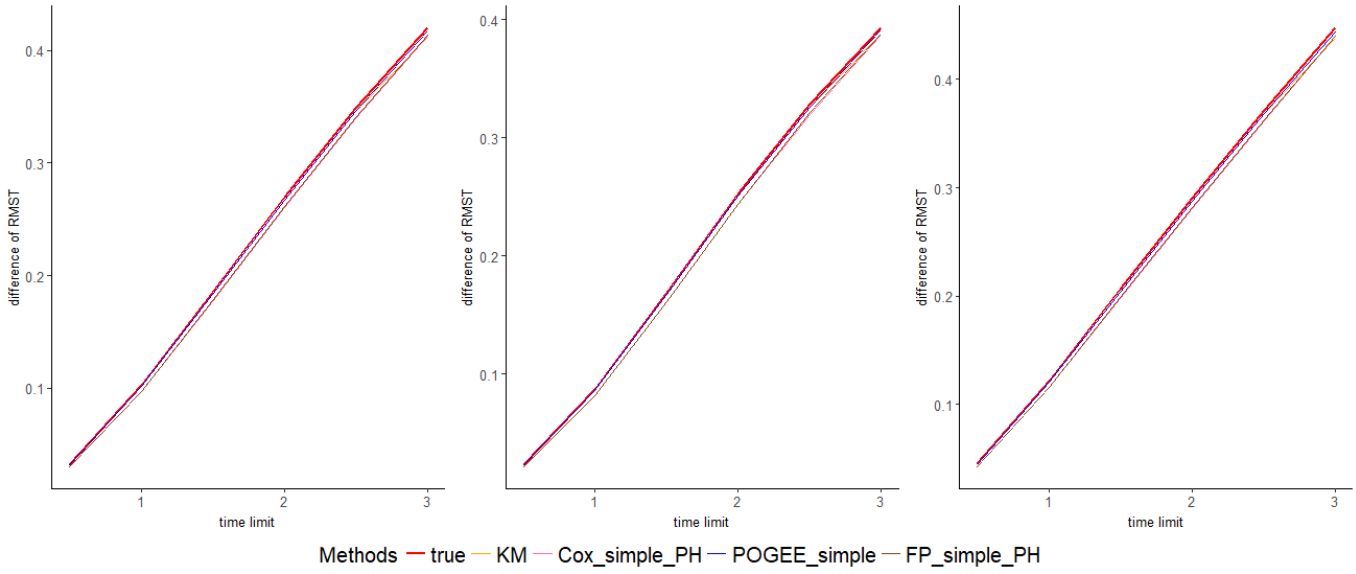


Figure 3.1: Methods performance estimating average RMST difference for a population under PH datasets with overall 25% exponential censoring. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

When the PH assumption is violated, however, the change in area between survival curves is nonmonotonic because of the crossing of the survival curves. Thus, the RMST difference is nonmonotonic. See Figure 3.2

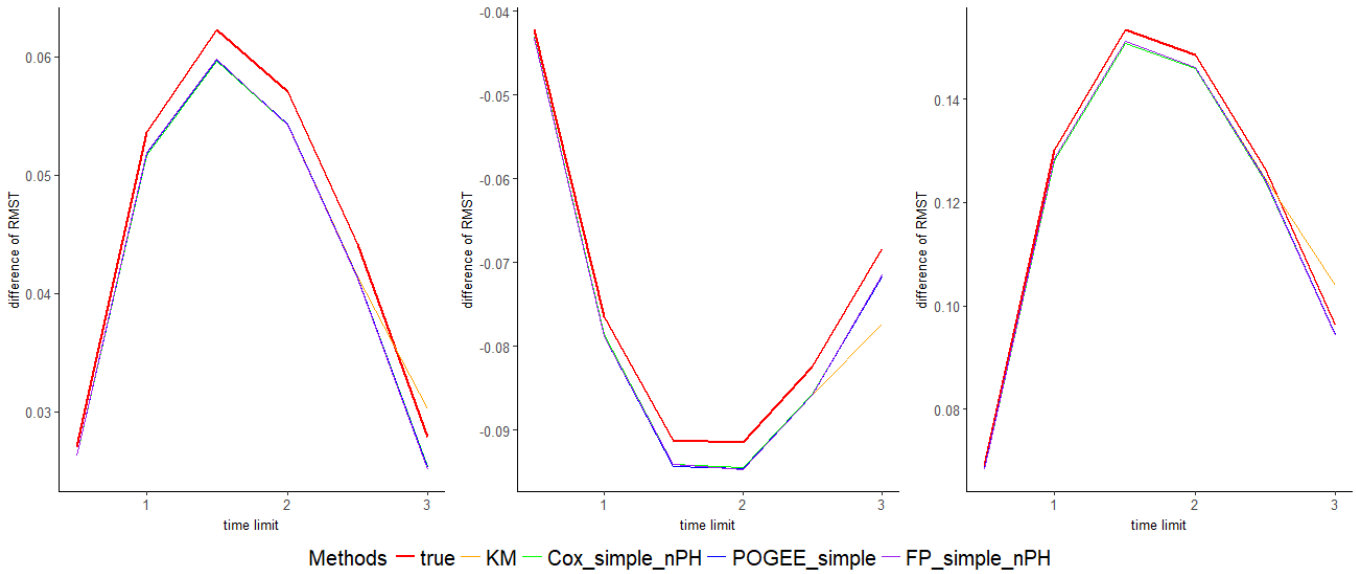


Figure 3.2: Methods performance estimating average RMST difference for the population under non-PH datasets with overall 25% exponential censoring . Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

If testing of the PH assumption is ignored, investigators may accidentally use methods that require fulfillment of the PH assumption to analyze treatment effects that may make their results invalid. The basic Cox model and flexible parametric model are both based on the proportional hazards assumption. Where the assumption is violated, the Cox model can be stratified. The flexible parametric model can also be generalized by the addition of more parameters to the model.

As shown in Figure 3.3, the two dash lines which present results of models holding PH assumption failed to detect the nonmonotonic change. The results for these two PH based models may get closer to the theoretical value than other models at some time points (e.g. where the time limit was set to 3 in the middle panel of Figure 3.3), but the results are still not generally acceptable. The Kaplan-Meier method and POGEE method are free from the PH assumption and offer promising results for PH and non-PH datasets. For the Cox model and the flexible parametric model, however, the PH assumption plays a significant role in applicability. The verification of PH assumptions is thus crucial when these two models are considered.

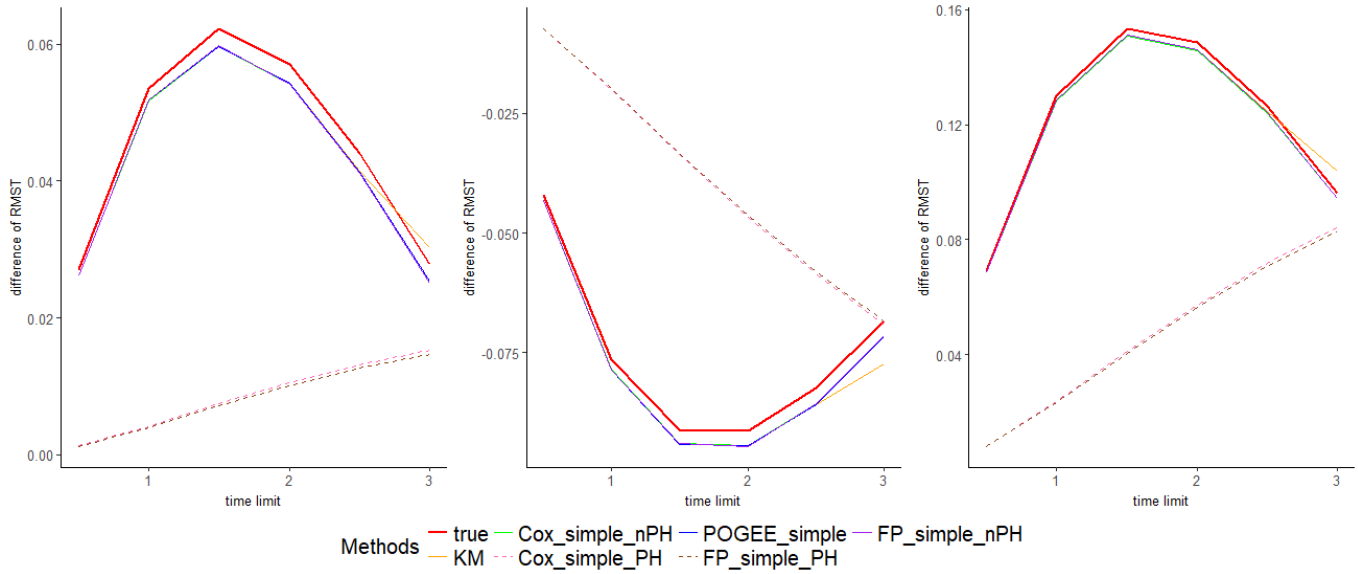


Figure 3.3: Methods performance estimating average RMST difference for a population in non-PH datasets with overall 25% exponential censoring. With methods which assume PH assumption. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

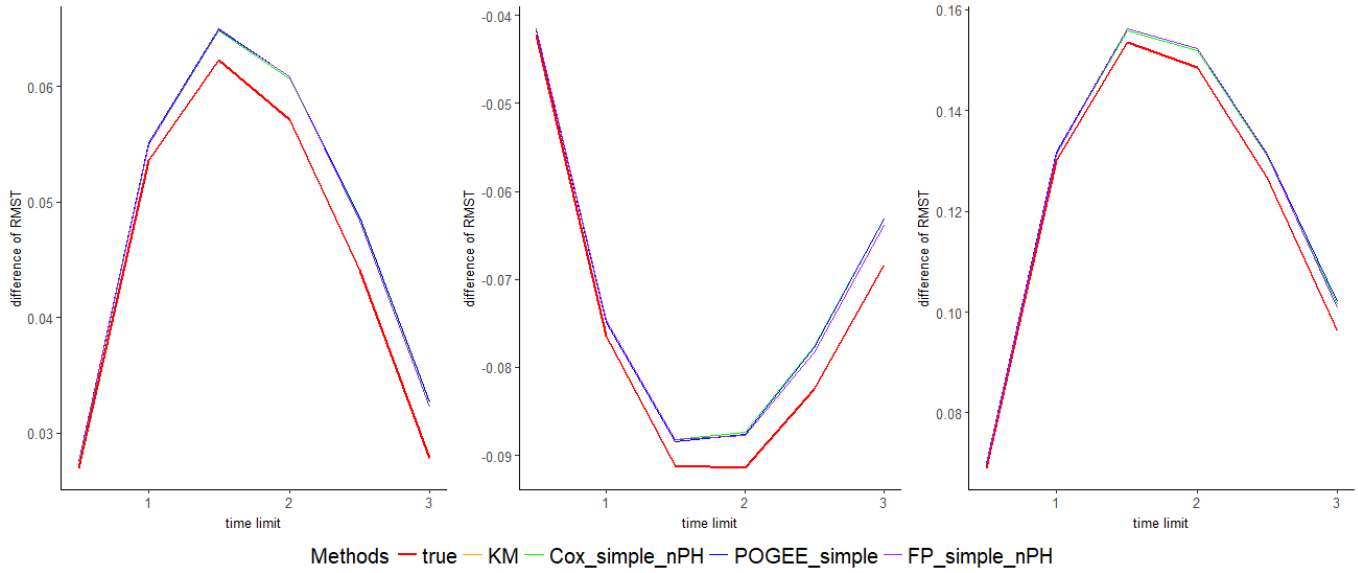


Figure 3.4: Methods performance estimating average RMST difference for a population in non-PH datasets without censoring. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

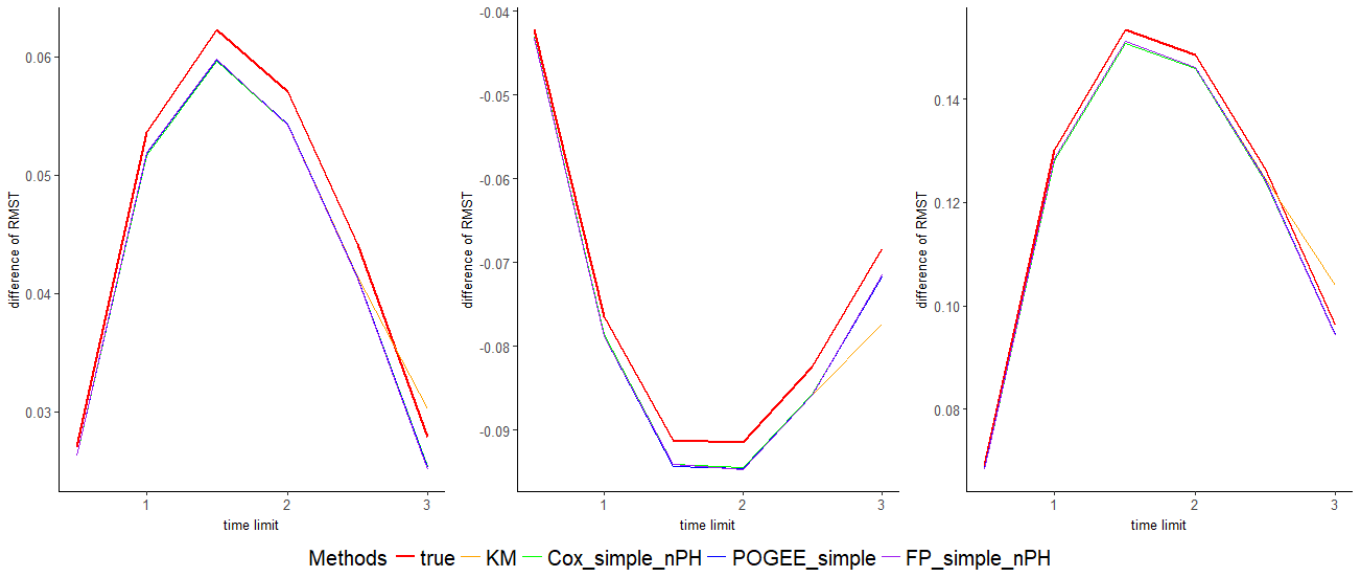


Figure 3.5: Methods performance estimating average RMST difference for a population in non-PH datasets with overall 25% exponential censoring. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

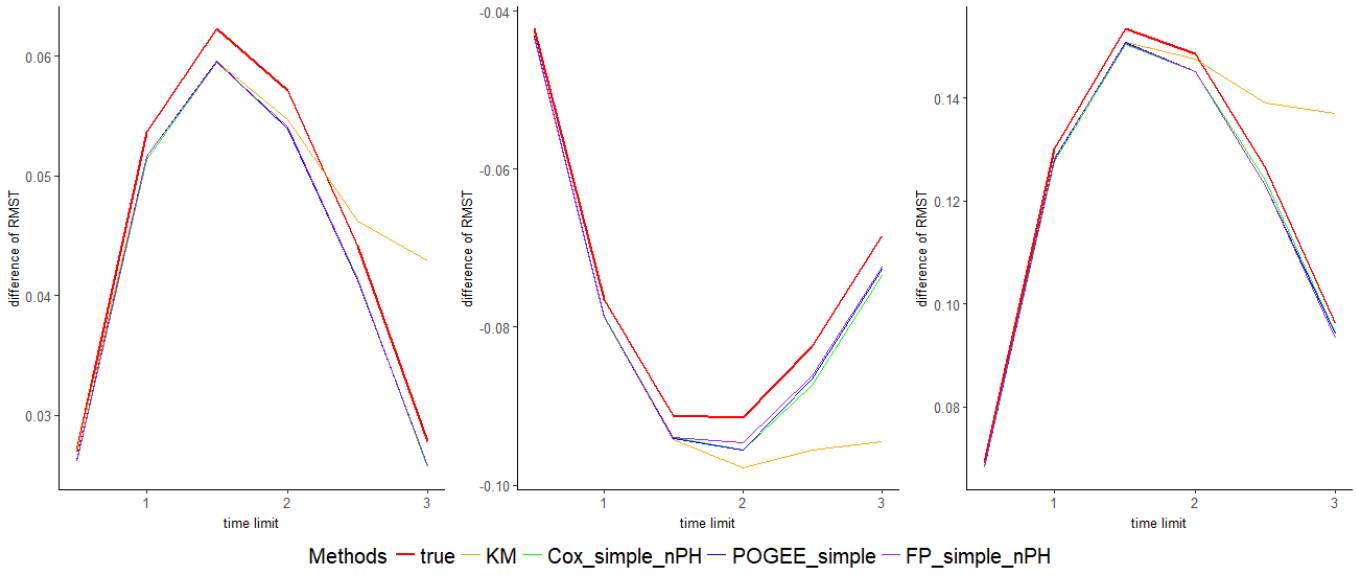


Figure 3.6: Methods performance estimating average RMST difference for a population in non-PH datasets with overall 50% exponential censoring. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

3.1.2 Influence of censoring proportion

Figures 3.4, 3.5 and 3.6 show the estimation results for different proportions of censoring. When no censoring occurs, the estimation results for all four methods are slightly larger than the theoretical values. When censoring does occur, the estimation results for all four methods are smaller than the theoretical values. The non-parametric Kaplan-Meier method is particularly significantly influenced by the censoring proportion under higher time limits, while the other three methods perform similarly under all time conditions. As a result, the other three methods are preferred to the simple Kaplan-Meier at time points in the right-hand tail. The standard deviation results also show poor performance for the Kaplan-Meier at later time limits (Figure 3.7).

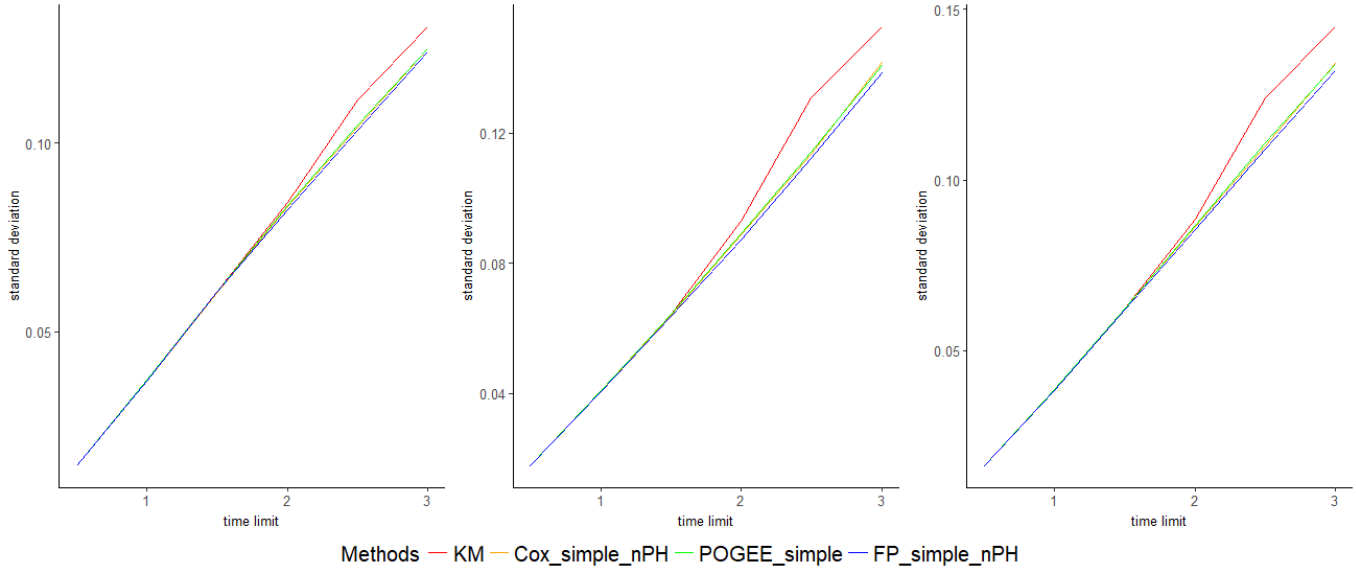


Figure 3.7: Standard deviation of methods estimating average RMST difference for a population in non-PH datasets with overall 50% exponential censoring. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

The RMST difference is a statistic that offers flexibility in terms of the time range. Although the simple Kaplan-Meier method performs less well in the right-hand time dimension, its estimations remain good for earlier time points. In Figure 3.5, the Kaplan-Meier method is seen to become deflected after 2.5 years when the overall censoring proportion is 25%. When the censoring proportion is set to 50% (Figure 3.6), the Kaplan-Meier results depart earlier but remain good prior to that departure. Choosing a smaller time limit, τ, also makes the overall censoring proportion in time range $[0, \tau]$ smaller than the overall censoring proportion of $[0, 3]$. By choosing the τ carefully, the simple Kaplan-Meier method can thus still be considered for estimating the RMST difference between treatment

group and control group for average survival.

Another way to improve the performance of the Kaplan-Meier with a time limit close to the trial duration is to increase the sample size. More subjects experience the event of interest, resulting in better survival function estimation.

3.1.3 Influence of censoring distribution

Figure 3.5 and Figure 3.8 show the estimation results under two different censoring distributions with the same overall censoring proportions. In all three situations, the estimation results with uniform censoring proportions are better than estimations with exponential censoring.

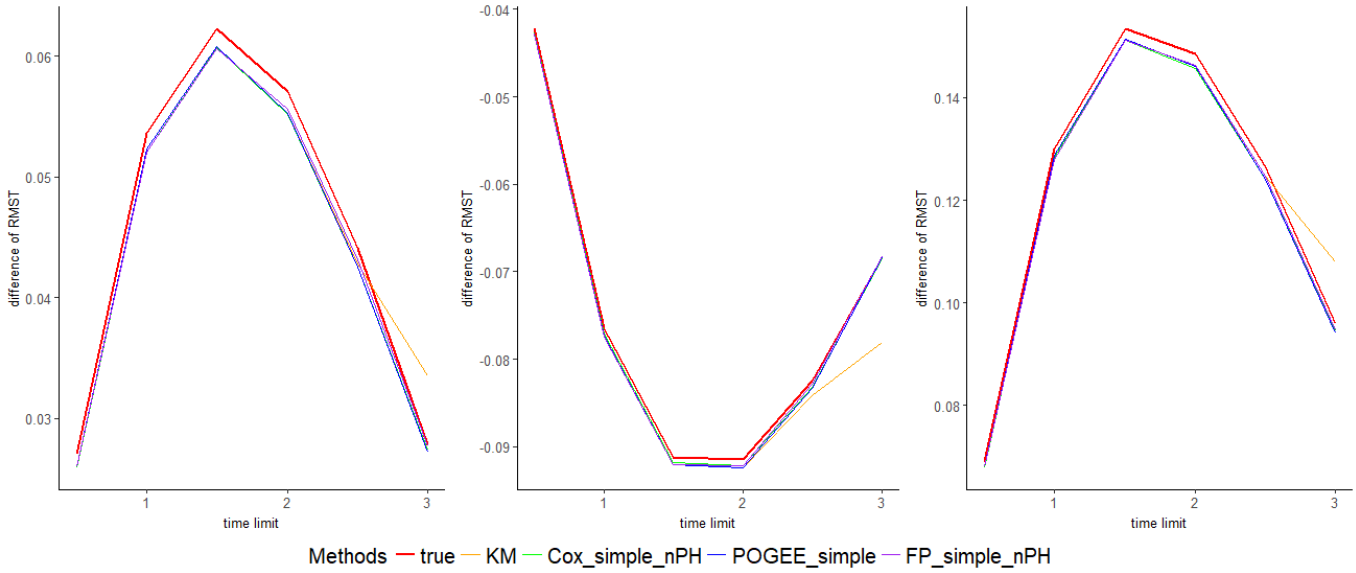


Figure 3.8: Methods performance estimating average RMST difference for a population in non-PH datasets with overall 25% uniform censoring. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

3.1.4 Influence of baseline hazard shape

Under non-PH situations, the varying combinations of baseline hazard shapes for treatment groups will determine the sign and variation tendencies of the RMST differences (Figures 3.4, 3.5, and 3.6). The RMST difference with constant hazard for the control group and increasing hazard for the treatment group increase to reach a peak around $t = 1.5$ and decrease afterward. The combination of constant hazard for the control group

and decreasing hazard for the treatment group causes the RMST difference to decrease firstly, then increase. The RMST difference with decreasing hazard for the control group and increasing hazard for the treatment group increases to reach a peak around $t = 1.5$ and then decreases, but a bigger difference is seen compared with the constant, increasing combination. On reversing the hazard shape for treatment groups in each combination, the direction of change and sign of RMST difference estimation is also reversed.

As mentioned in the previous subsection, and as shown in Figures 3.5 and 3.6, Kaplan-Meier methods demonstrated a departed result at later time points with censoring. Comparing the three different combinations, the departure distance in the left panel is the smallest and the performance at time point 2.5 is better than in the other two combinations. While the departure direction in the middle panel becomes smaller than the theoretical value, the estimation of RMST difference using the Kaplan-Meier method in other two panels becomes bigger than the true value as the trial end time approaches. The deviation distance is largest in the right-hand panel.

3.1.5 Estimation for a population with different ages and genders

The simple Kaplan-Meier method is a non-parametric method for estimating the average survival curve for a population comprised of individuals of different ages and genders. The other three methods also estimate the RMST for average survival when only the treatment covariate is included in the (semi-)parametric model. As seen in Figures 3.4, 3.5, and 3.6, all four methods' performance are very similar without censoring. As the pseudo-observation of RMST is based on the Kaplan-Meier estimation, the estimation of this method will always get exactly the same result as the Kaplan-Meier when no censoring occurs. When censoring does occur, the Kaplan-Meier shows a very poor performance at later time points, while the other three methods continue to perform well.

3.1.6 Estimation for a subject with given ages and gender

For RMST difference estimation, it is also possible to analyze the results for a subject with given age and gender. This work thus estimated the results for a male with age set to 50.

The Kaplan-Meier method no longer fits in this situation, so a comparison was conducted only among the three remaining methods, adjusted for all treatment, age, and gender covariates. As shown in Figure 3.9, the Cox model and the Flexible parametric model perform well. The POGEE method, however, was also seen to be not adequate for

this situation. The pseudo-observation of RMST used in the parametric linear model is still a non-parametric estimation based on Kaplan-Meier, and in the linear model fitting step for the 1,000 simulated datasets, most of the estimated treatment coefficients are not significant, and the coefficients for age and gender do not influence the estimation results sufficiently. Comparing the estimation results for POGEE in Figure 3.9 (estimate for 50-year old male) and Figure 3.5 (estimate for a population), the estimation results are nearly the same.

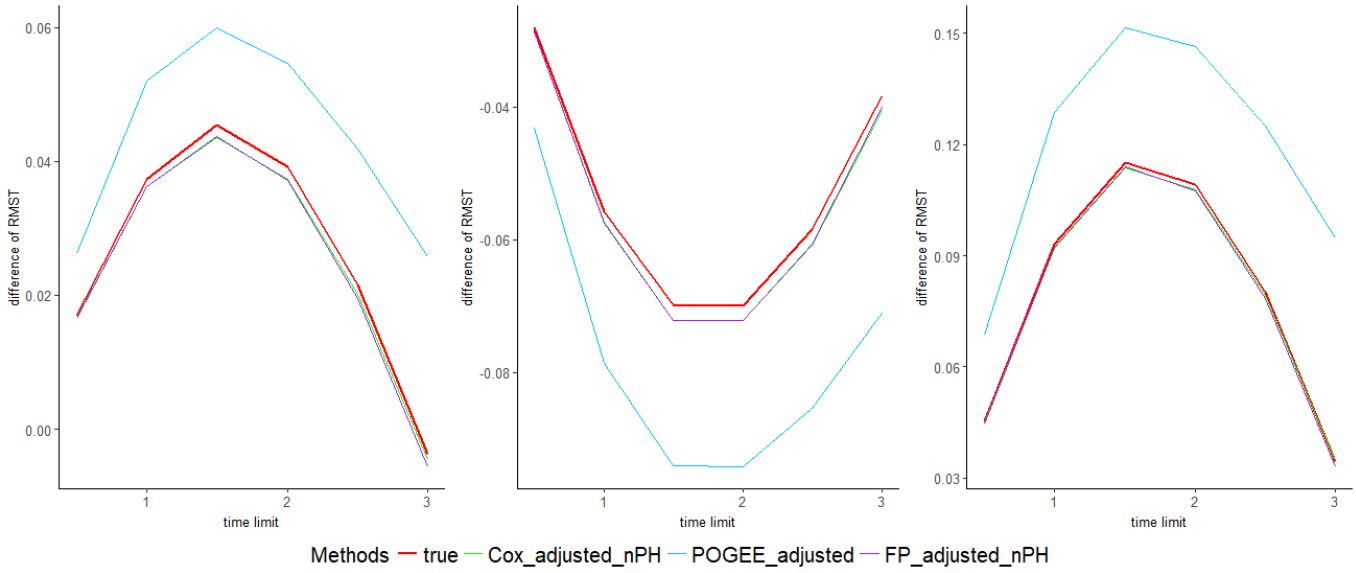


Figure 3.9: Methods performance estimating RMST difference for male with age 50 under non-PH datasets with overall 25% exponential censoring. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

3.2 proportional hazards situation

Non-proportional hazards situation is the main focus of this work; however, a more comprehensive understanding can be developed by also including the performance of methods under PH situations. As the RMST statistic is expected to act as a general measurement that suits both PH and non-PH situations.

As shown in Figure 3.1, the change of estimation of RMST difference between groups will always be monotonic.

3.2.1 Influence of censoring proportion

When no censoring happens (Figure 3.10), the line present results from the Kaplan-Meier method nearly overlap with the line shown theoretical values; as the POGEE method is built on Kaplan-Meier estimation, it also achieves better performance than the other two methods. Similar to the non-PH situation, here under PH situation, the estimation achieved using the Kaplan-Meier method also begins to display a bias towards the later time points when censoring happens (Figures 3.11 and 3.12). The degree of deviation gets larger as the proportion of censoring increases. Figures 3.10, 3.11, and 3.12 present the differences between methods more clearly where the time range in the x-axis is set to $[2.5, 3]$. In the time range $[0, 2]$, the difference is smaller.

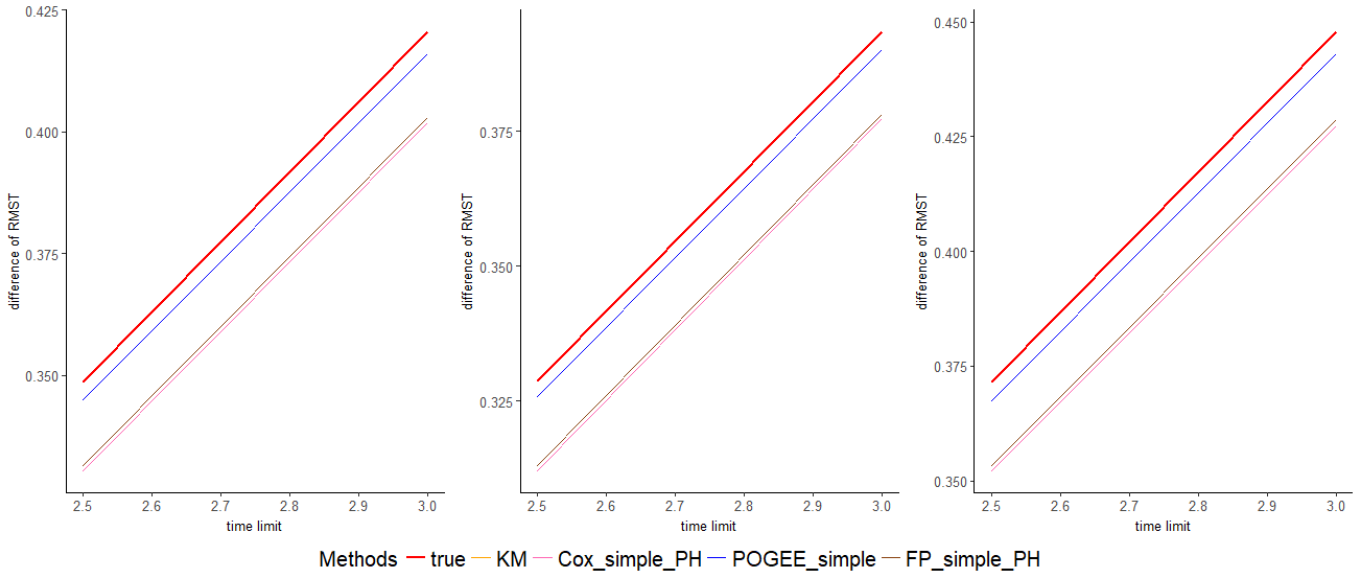


Figure 3.10: Methods performance estimating average RMST difference for a population in PH datasets without censoring. Limited time range $[2.5, 3]$. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

3.2.2 Influence of censoring distribution

Figures 3.11 and 3.13 show estimation results under different censoring distributions with the same overall censoring proportions. Similar to the results in the non-PH section, all methods perform better and achieve estimations closer to the theoretical value with uniform censoring than with exponential censoring.

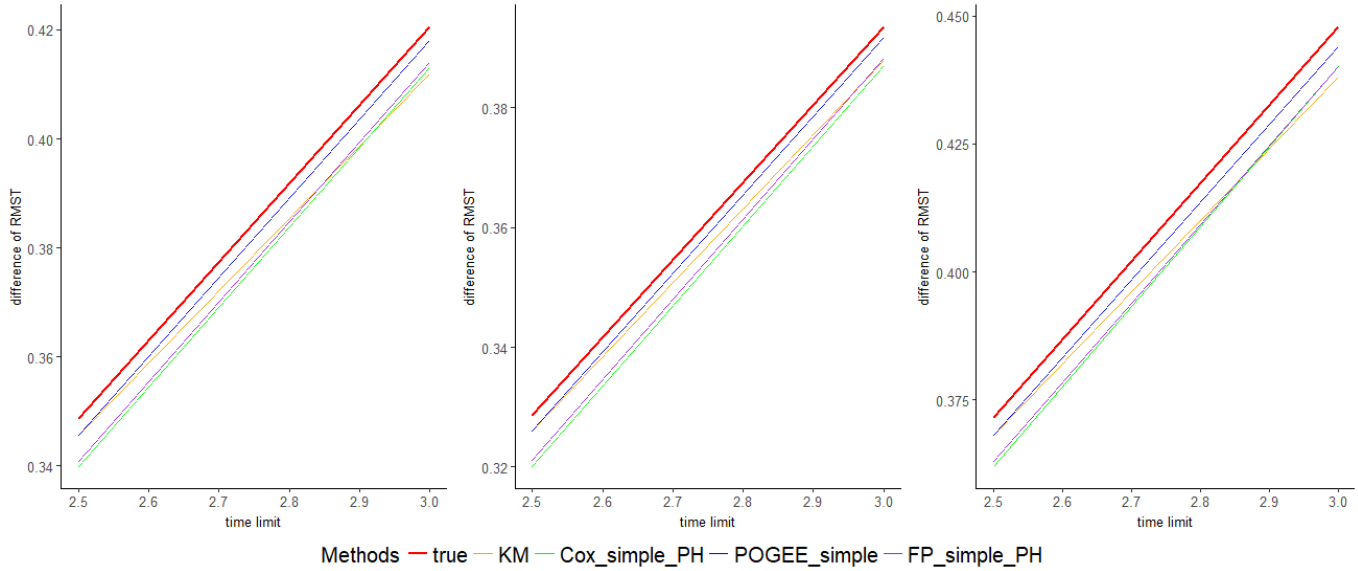


Figure 3.11: Methods performance estimating average RMST difference for a population in PH datasets with overall 25% exponential censoring. Limited time range $[2.5, 3]$. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

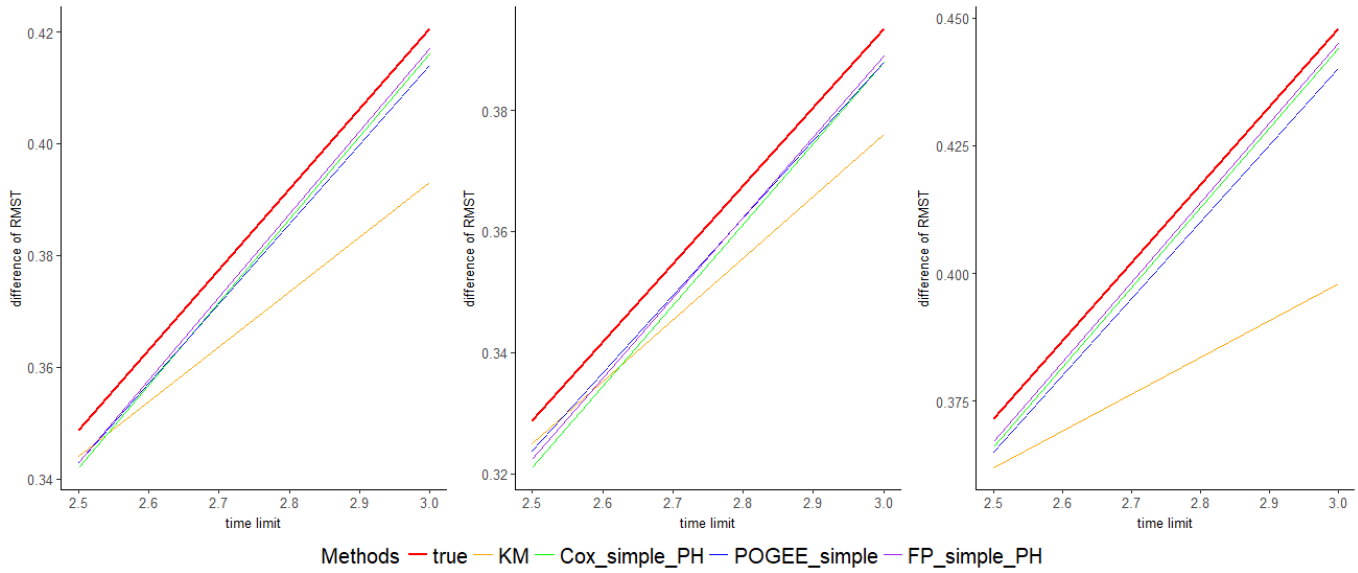


Figure 3.12: Methods performance estimating average RMST difference for a population in PH datasets with overall 50% exponential censoring. Limited time range $[2.5, 3]$. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

3.2.3 Influence of baseline hazard shape

Compared to the results in the non-PH section, the changes in the RMST difference results for PH populations all monotonically increase with a positive sign. The shape and

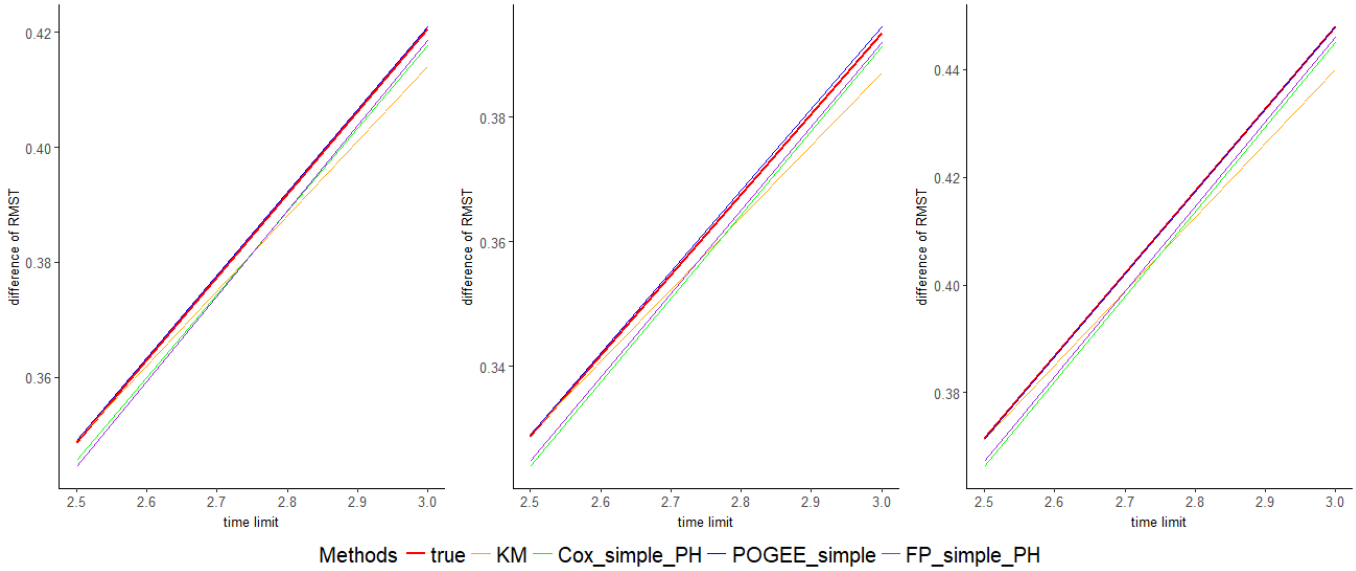


Figure 3.13: Methods performance estimating average RMST difference for a population in PH datasets with overall 25% uniform censoring. Limited time range $[2.5, 3]$. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

increase rates are similar among all three situations. The y-axis in Figure 3.14 shows that the decreasing baseline hazard situation (Right panel) has the maximum RMST difference result at each time limit, with the constant baseline hazard situation next, and the increasing baseline hazard situation (Middle panel) showing the smallest result. Generally, the baseline hazard shape does not have a significant influence on the estimation result. For the Kaplan-Meier method, a decreasing baseline hazard situation results in a higher deviation distance in the later time range (Figure 3.12).

3.2.4 Estimation for a population with different ages and genders

Figures 3.10 to 3.14 show the estimation results for average RMST differences in populations with varying ages and genders which suggest that, under PH situations, all four methods perform similarly. The standard deviations over 1,000 samples for all methods except the Kaplan-Meier are close to each other, while the Kaplan-Meier displays higher standard deviations at later time points (Figure 3.15).

As previously discussed, Kaplan-Meier method lacks the ability to adjust for additional covariates. The other three methods examined can incorporate, but when the censoring proportion is small and the average RMST difference for the population are estimated, these three methods do not provide better estimations than the Kaplan-Meier. For the Cox and flexible parametric methods, estimating the RMST difference adjusted for average age

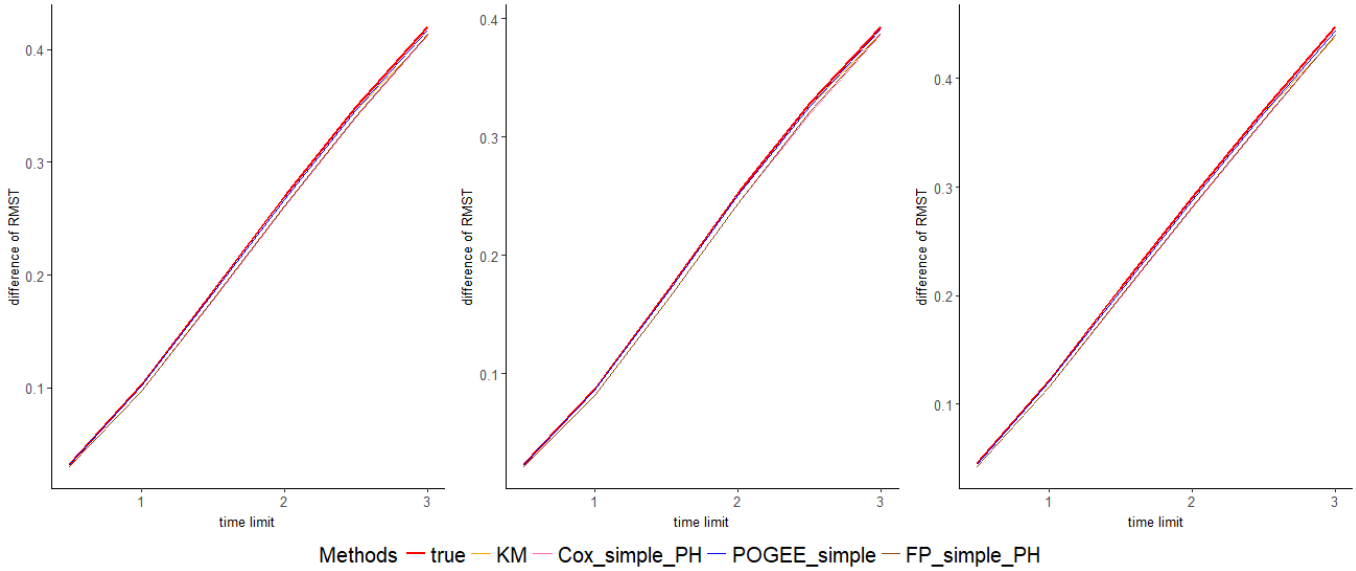


Figure 3.14: Methods performance estimating average RMST difference for a population in PH datasets with overall 25% exponential censoring. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

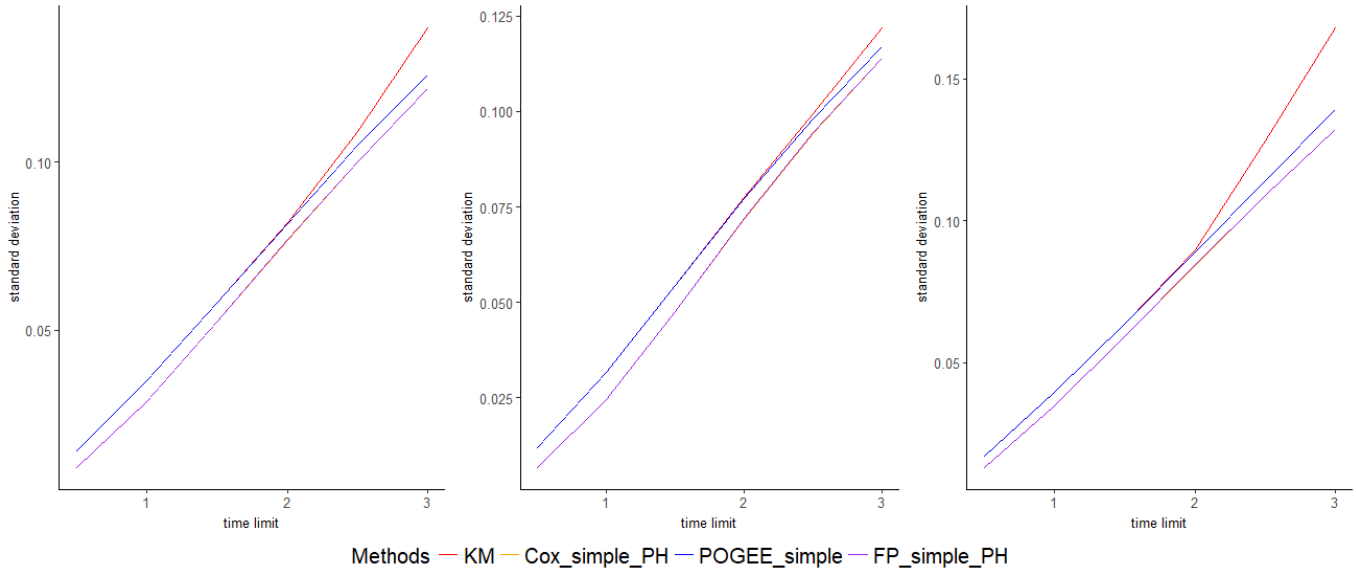


Figure 3.15: Standard deviation of methods estimating average RMST difference estimates for a population in PH datasets with overall 50% exponential censoring. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

and average gender is not identical to estimating average RMST difference by including treatment as the only covariate (Figure 3.16), the two dashed lines presented adjusted Cox model and adjusted flexible parametric model obtain bigger results compare to the theoretical values. And they are more further away from the theoretical values than the simple model estimations (Figure 3.16). For POGEE, adjusting for age and gender does

not influence the results much. The Kaplan-Meier method can reach an estimation closer to the theoretical value than any of the other three methods at early time points when the censoring proportion is between [0, 0.25] (Figure 3.10 and 3.11).

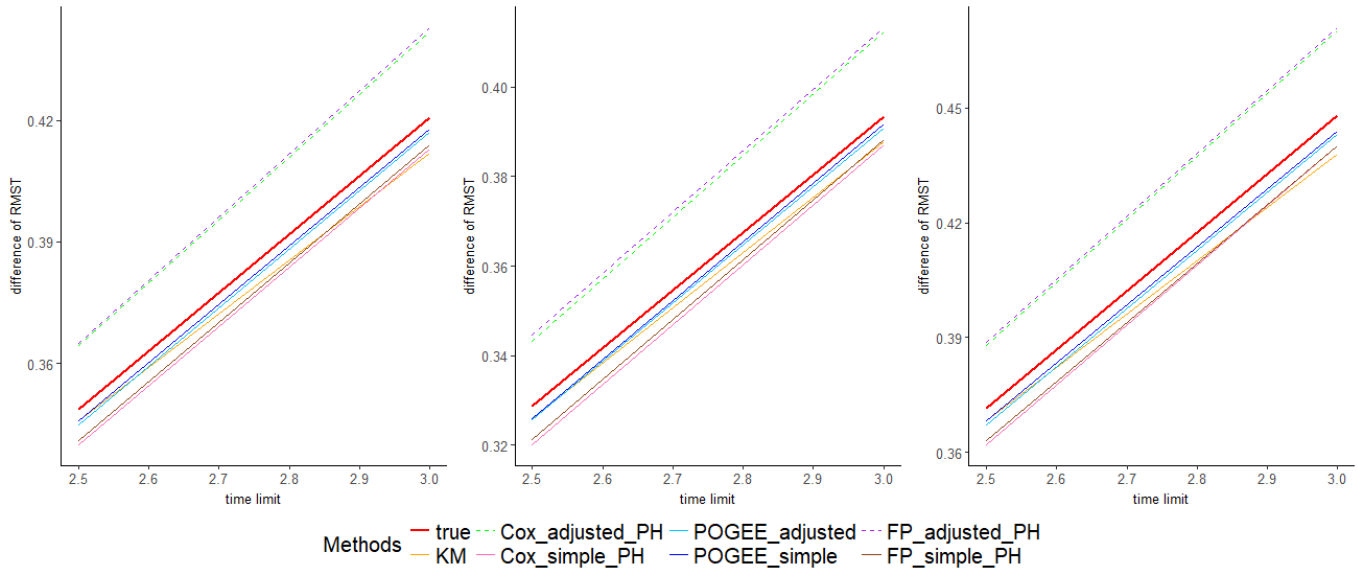


Figure 3.16: Methods performance estimating average RMST difference for a population in PH datasets with overall 25% exponential censoring. Limited time range $[2.5, 3]$. Include models adjusted for average age and average gender. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

3.2.5 Estimation for a subject with given ages and genders

The Kaplan-Meier methods no longer apply in this situation. For the other three methods, the POGEE method performs worse than the other two methods. For all three baseline hazard shapes, the performance of all three methods is similar (Figure 3.17). The standard deviation of POGEE is also larger (Figure 3.18).

3.3 Computation time

The Kaplan-Meier method, stratified Cox model method, and flexible parametric method all follow the same procedure for estimating RMST by first estimating the survival function and then calculating the area under the survival curve. As shown in Table 3.1, the Kaplan-Meier method and stratified Cox model have similar computation times in both steps. They use the same function to estimate RMST based on stepped survival curve. The function adds the areas of rectangles formed by the stepped survival curve to calculate RMST results rapidly. The estimated survival curve using the flexible parametric

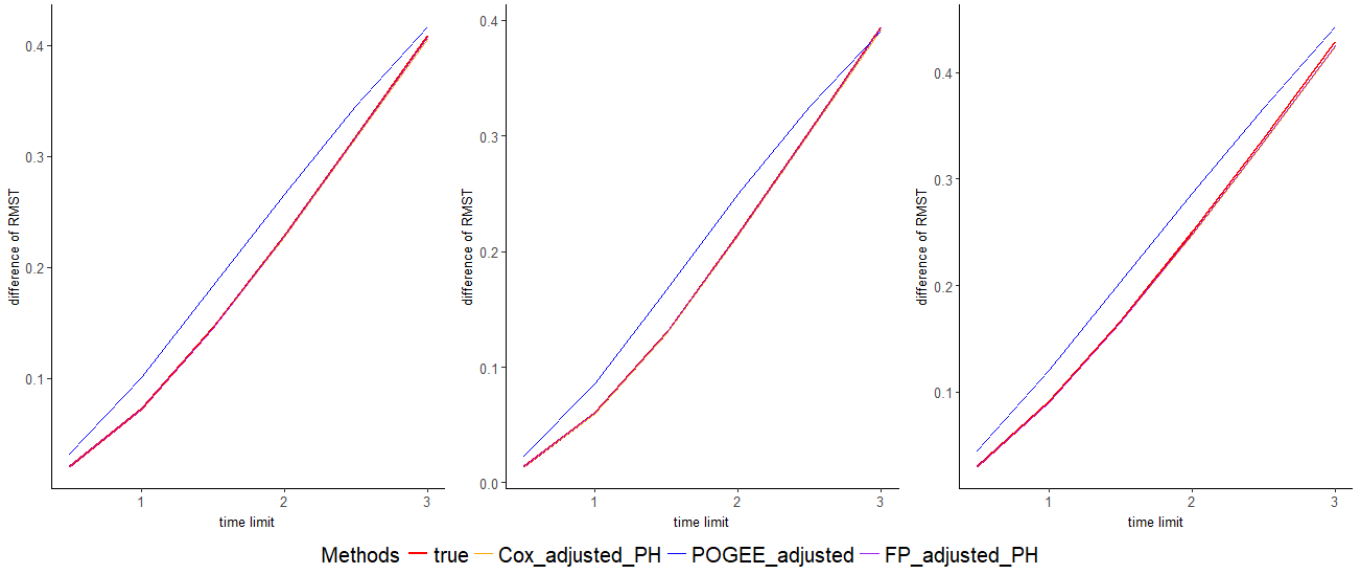


Figure 3.17: Methods performance estimating RMST difference for male with age 50 in PH datasets with overall 25% exponential censoring. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

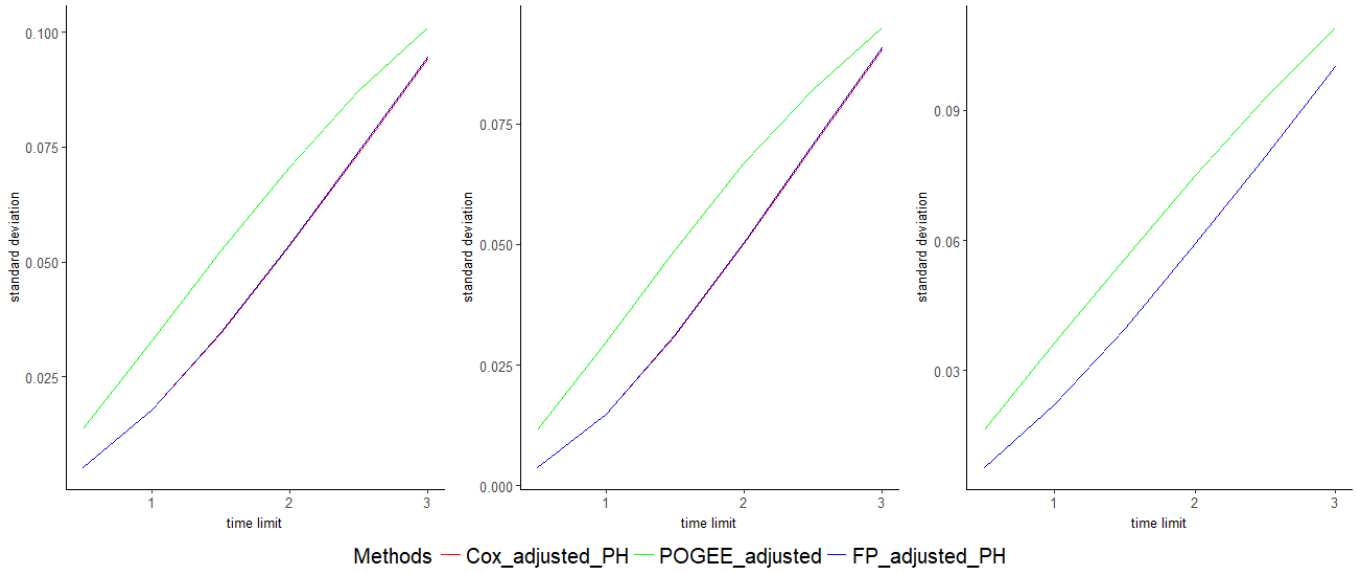


Figure 3.18: Standard deviation of methods estimating RMST difference for male with age 50 in PH datasets with overall 25% exponential censoring. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

method cannot be described by a simple step function, and a 15-point Gauss-Kronrod quadrature method is thus used to estimate the integral of the survival function. This integral calculating step is more time-intensive than that used in the other two methods.

In the POGEE method, the pseudo-observations of RMST must be computed first and then a generalized linear model was built treating pseudo-observation of RMST as

response variable and covariates as independent variables. After obtaining the coefficients' estimation using GEE, a prediction of RMST can be made using the given values of the covariates. Regarding time taken to get the final estimate, the result here was 104.22(seconds). When compared with the total time for each of the other three methods, this computation time is significantly shorter.

Table 3.1: The computation time for four methods.

	Estimating survival (second)	Estimating RMST (second)	Total time (second)
Kaplan-Meier	140.18	0.16	140.34
Cox model	150.62	0.19	150.81
Flexible parametric model	403.83	2317.56	2721.39
POGEE	not applied	not applied	104.22

All methods are running under a non-PH situation with decreasing baseline hazard for control group and increasing baseline hazard for treatment group. Censoring as 25% unifrom censoring. Time limit is 1.5 years. Computation time is the total time for 1000 replications.

3.4 Overall summary

No method appears significantly better than other methods in all simulated scenarios.

3.4.1 Performance of methods under non-proportional hazards situations

The Kaplan-Meier method is only suitable for estimating average RMST differences for a population; however, under low censoring proportions (0% to 25%) and early time limit (smaller than 2.5), the performance of the Kaplan-Meier method is on par with other methods. Considering the simplicity of Kaplan-Meier method and the fact that it is free from the assumption (other than the independence of censoring), it should thus be preferred where these situations hold. This recommendation is heavily dependent on the choice of the time limit τ, however. The estimation of Kaplan-Meier methods with time limits near the end trial time or last observation time is less desirable. Compared to the other methods, higher censoring proportions also negatively influence the performance of this method significantly.

In this thesis, the Cox model was stratified by the treatment covariate when the hazard ratio between treatment groups was not proportional. This generalization pattern worked well and gave promising results most of the time in the simulation scenarios. The flexible parametric model was generalized for non-proportional situations in a different way and it performed similarly to the stratified Cox model most of the time. It achieved better estimation than the Cox model under some combinations of baseline hazards with high

censoring proportions (Figure 3.6, middle panel), but the differences were very small. The main shortcoming of the flexible parametric method is the time-consuming integration calculating step in R software. In other software (e.g., SAS), the computation time may be better. When estimating RMST differences for a subject with given age and gender, these two methods are the most appropriate methods.

The pseudo-observation combined with linear model and GEE method (POGEE) has a totally different estimation procedure from that used by the other three methods. When estimating average RMST difference for a population, this method is also recommended. Compared with Kaplan-Meier, it gives promising estimations at larger censoring proportions and later time limits. Compared to the Cox model and flexible parametric models, it requires no special extensions for non-proportional situations and it is less time-consuming. It also gives regression results on how covariates influence the pseudo-observation of RMST. However, in our simulation study, the regression results do not influence the RMST estimation significantly. Considering the calculation of pseudo-observation is based on Kaplan-Meier and its performance in the simulation study, this POGEE method is not suitable for estimating RMST difference for a subject with given age and gender.

3.4.2 Performance of methods under proportional hazards situations

Under low censoring proportion (0% to 25%) and early time limit (smaller than 2.5), the performance of Kaplan-Meier method is the best. Its performance under larger censoring proportions and larger time limits are similar to its performance in non-PH situations. Compared with other methods, the baseline hazard shape and censoring proportion have a significant influence on the performance of Kaplan-Meier estimations. A decreasing baseline hazard shape and a high proportion of censoring will result in the most biased estimations at later time points (Figure 3.12).

For the Cox and flexible parametric models, the proportional hazards assumption requires no extensions to the models, unlike non-PH situations. Both methods give good estimations, but the standard deviations for large time limits are large (see Figure 3.15 and Appendix). The computation time comparison results in Table 3.1 are given under a non-PH situation. When estimating for PH datasets, the time-consuming issues of the flexible parametric method is also serious.

The POGEE method has a larger standard deviation than the Cox and flexible parametric models under high censoring proportions (Figure 3.15). Similar to non-PH scenarios, the POGEE method performs worse than Cox and FP models estimating RMST

difference for a 50 years old male subject (Figure 3.17). When estimating the average RMST difference for a population, its performance is on the same level as other three methods and perform best under 25% exponential censoring (Figure 3.11).

Chapter 4

Discussion

4.1 Innovation and conclusion

This thesis described four different methods as proposed in previous articles for estimating RMST: the Kaplan-Meier method, Cox model method, Flexible parametric model method, and pseudo-observation combined with a linear model and GEE (POGEE). A relatively comprehensive comparison of performance between these four different methods was conducted. Royston and Parmar [**1**] made comparisons between the Kaplan-Meier, Flexible parametric model, and POGEE method by using three real datasets. Compared to that work, this thesis includes more methods, and the simulation studies provide a more general view of how these methods perform under different data types, censoring proportions, time ranges, and target populations.

Based on the Results chapter, we can conclude the performance of methods and offer some suggestions. The Kaplan-Meier method gives acceptable performance regarding estimating average RMST difference for a population with different ages and genders under early time limits and scenarios with small censoring proportions. The requirements necessary to generate promising results using the Kaplan-Meier method are, however, greater than for other methods; nevertheless, its simplicity makes it preferable when those requirements are met. When the datasets do not meet these requirements, the POGEE method performs better than the Kaplan-Meier method in terms of estimating average RMST difference for a population. The other two models also give promising results, but they are more time-consuming. When estimating RMST difference for a subject of specific age and gender, the Cox model and Flexible parametric methods are the only two methods that are suitable. And they perform similarly at all scenarios. Considering the computation time, Cox model is then a better choice than the flexible parametric model

since the flexible parametric model is too time-consuming.

4.2 Limitations and future work

During the conduction of the simulation studies, some limitations emerged. The first limitation is about censoring. We introduce two aspects concern censoring to compare the performance between methods: censoring distribution and censoring proportion. Within the simulated datasets used in this work, the subjects were forced to have events mostly before the trial end time, set to 3. Based on this, the censoring times were added. However, an overall 50% proportion of uniform censoring was hard to reach under the simulated event times, although the exponential censoring was more flexible and allowed a higher censoring proportion to be reached. A comparison between the two censoring distributions at the 50% level was thus not possible. Another limitation concerned the flexible parametric method. All of the results used in this thesis were computed in R, and the long computation time for the flexible parametric model was impacted by the built-in $integrate$ function and $flexsurv$ package. Using different software (e.g. SAS) may thus improve the computation time. The other way to decrease the computation time is setting the starting value for the parameters in the flexible parametric model [**8**].

The results presented in this thesis do inspire possible further work. The first would be to seek an explanation of the intermediate results from the POGEE method. In the simulation studies, the regression results from each sample were quite unstable and whether the coefficients for covariates significantly influence the estimation is a concern. When using this method for a real dataset, additional model checking procedures would be required. An analysis of the scatterplot was shown in [**8**].

Another point concerns the Cox model. In the simulation studies, the performance of the stratified Cox model was quite good under non-proportional hazards situations. This may be connected to the method used in this thesis for building non-PH datasets. Under some situations, such as small risk set size, this stratified model may not be stable [**1**]. Although the stratified Cox model sets different baselines for different treatment groups and it creates a stepped function for the survival function estimation, it gave quite promising estimation results under higher censoring proportions and later time limits where the Kaplan-Meier became biased. According to the help page for $survfit(coxph)$, the survival result estimation from a stratified Cox model adjusted for no other covariates will follow the exponential of Nelsons cumulative hazard estimates. The survival results based on Nelsons cumulative hazard estimates again forms a step function which is higher than the Kaplan-Meier estimated stepped survival curve. In the simulation settings, the results for

the stratified Cox model were thus seen to be acceptable when calculating the difference of RMST between two arms. The deeper reason for this difference in performance between the two stepped estimation results may, however, require more investigation. One possible reason is connected to the difference between the Kaplan-Meier estimator and Nelson-Aalen estimator.

An important point about estimating RMST is the choice of time limit τ. In this thesis, RMST difference was estimated in round numbers that were deemed more suitable for explaining the results in the form 'taking the treatment will increase the life expectancy by δ years in the next τ years compared with not taking treatment'. Other than selecting τ to be a round number, the choice of time limit τ may also be connected to the observation situation. In the comparison section of Royston and Parmar [**1**], the whole observed follow-up time was deemed to be the period of interest. In the current simulation scenarios, the upper limit for the whole observed follow-up time lies almost completely beyond time point 2.5, and thus the Kaplan-Meier method is not suitable. The functions used in this thesis are very flexible in terms of selecting a time limit, but in some other software types, such as *SAS*, the built-in choices of τ are limited. Considering this type of time limit choice, further investigation of the differences between choosing the time limit as the last event time, last observed time, and a time between the previous two time marks may be needed.

The simulation study in this thesis only includes right censoring into consideration. In real cases among oncology area or other clinical trials where a non-PH situation occurs, left truncation could also happen. Further work on left truncation is thus required.

Due to the time limitations of this thesis, the coverage of confidence intervals for different methods was not compared. The bootstrap method for flexible parametric model methods is too time-consuming when seeking confidence interval results for large replications. The code that could be used to calculate confidence intervals for all four methods is presented in Appendix A.

Bibliography

[1] Royston, P., & Parmar, M. K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine, 30*(19), 2409-2421. doi:10.1002/sim.4274

[2] AHern, R. P. (2016). Restricted Mean Survival Time: An Obligatory End Point for Time-to-Event Analysis in Cancer Trials? *Journal of Clinical Oncology, 34*(28), 3474-3476. doi:10.1200/jco.2016.67.8045

[3] Andersen, P. K., Hansen, M. G., & Klein, J. P. (2004). Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations. *Lifetime Data Analysis, 10*(4), 335-350. doi:10.1007/s10985-004-4771-0

[4] Royston, P., & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine, 21*(15), 2175-2197. doi:10.1002/sim.1203

[5] Irwin, J. O. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene, 47*(02), 188-189. doi:10.1017/s0022172400014443

[6] Kalbfleisch, J. D., & Prentice, R. L. (1973). Marginal Likelihoods Based on Coxs Regression and Life Model. *Biometrika, 60*(2), 267. doi:10.2307/2334538

[7] Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine, 19*(9), 1141-1164. doi:10.1002/(sici)1097-0258(20000515)19:93.0.co;2-f

[8] Andersen, P. K., & Perme, M. P. (2009). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research, 19*(1), 71-99. doi:10.1177/0962280209105020

[9] Kuonen, D. (2003). Numerical Integration inS-PLUSorR: A Survey. *Journal of Statistical Software, 8*(13). doi:10.18637/jss.v008.i13

[10] Klein, J. P., & Moeschberger, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*. New York: Springer.

[11] Lee, C. H., Ning, J., & Shen, Y. (2017). Analysis of restricted mean survival time for length-biased data. *Biometrics, 74*(2), 575-583. doi:10.1111/biom.12772

[12] Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine, 24*(11), 1713-1723. doi:10.1002/sim.2059

Appendices

A Appendix : code for thesis

A.1 brief instruction about how to use codes below

There are all six functions for data generation. The name for functions indicates the data type (PH or nPH) and censoring situation (nothing or with 'cen'). For non-proportional hazards, total four parameters have to be indicated. 'lamb0' and 'alpha0' are parameters of Weibull distribution for the control group. 'lamb1' and 'alpha1' are parameters of Weibull distribution for the treatment group. For proportional hazards data, two parameters have to be indicated: 'lamb' and 'alpha' is parameters of Weibull distribution for the baseline group. When building data with censoring, the parameters for uniform distribution and exponential distribution is changeable.

All codes show below in each method part is designed for estimating results for a non-proportional hazards data without censoring for the simulation study in this thesis. Estimation for other data type and censoring situation is accessible by changing the data generation function name and Weibull distribution parameters accordingly.

To get the result for a simulation study, for each method, run functions under section 'Basic functions' first, then run codes under section 'Running results' after given Weibull parameters and time limit number. The codes are also adjustable for a real dataset, just delete the Weibull parameters and change the input as a dataset in 'Basic functions' part and ignore the replicate step in 'Running results' part.

The following packages have to be installed: *simsurv*, *survival*, *pseudo*, *geepack* and $flexsurv$. The plots are drawed using *ggplot*2 package.

A.2 code for data generation

Generating non-proportional hazards data without censoring:

```
#lamb0 and alpha0 is parameters of Weibull distribution for the control group
#lamb1 and alpha1 is parameters of Weibull distribution for treatment group
data_nPH <- function(lamb0, lamb1, alpha0, alpha1){
covs_t0 <- data.frame(id = 1:200, age = (round(runif(200, 20, 60))-40)/20,
gender = rbinom(200, 1, 0.5))

sim_data_nPH_t0 <- simsurv(dist = "weibull", lambdas = lamb0, gammas = alpha0,
x = covs_t0, betas = c(age = -0.5, gender = -0.5))

covs_t1 <- data.frame(id = 201:400, age = (round(runif(200, 20, 60))-40)/20,
 gender = rbinom(200, 1, 0.5))

sim_data_nPH_t1 <- simsurv(dist = "weibull", lambdas = lamb1, gammas = alpha1,
x = covs_t1, betas = c(age = -0.5, gender = -0.5))

#combine data
sim_data_nPH_t0 <- cbind(sim_data_nPH_t0, rep(0, 200), covs_t0[, 2:3])
names(sim_data_nPH_t0) <- c("id", "obs_tte", "eventYN", "treat", "age", "male")

sim_data_nPH_t1 <- cbind(sim_data_nPH_t1, rep(1, 200), covs_t1[, 2:3])
names(sim_data_nPH_t1) <- c("id", "obs_tte", "eventYN", "treat", "age", "male")

sim_data_nPH <- rbind(sim_data_nPH_t0, sim_data_nPH_t1)

return(sim_data_nPH)
}
```

Generating non-proportional hazards data with uniform censoring:

```
data_nPH_cen <- function(lamb0, lamb1, alpha0, alpha1){
  covs_t0 <- data.frame(id = 1:200, age = (round(runif(200, 20, 60))-40)/20,
  gender = rbinom(200, 1, 0.5))

sim_data_nPH_t0 <- simsurv(dist = "weibull", lambdas = lamb0, gammas = alpha0,
```

```
x = covs_t0, betas = c(age = -0.5, gender = -0.5))

covs_t1 <- data.frame(id = 201:400, age = (round(runif(200, 20, 60))-40)/20,
gender = rbinom(200, 1, 0.5))

sim_data_nPH_t1 <- simsurv(dist = "weibull", lambdas = lamb1, gammas = alpha1,
x = covs_t1, betas = c(age = -0.5, gender = -0.5))

#combine data
sim_data_nPH_t0 <- cbind(sim_data_nPH_t0, rep(0, 200), covs_t0[, 2:3])
names(sim_data_nPH_t0) <- c("id", "obs_tte", "eventYN", "treat", "age", "male")

sim_data_nPH_t1 <- cbind(sim_data_nPH_t1, rep(1, 200), covs_t1[, 2:3])
names(sim_data_nPH_t1) <- c("id", "obs_tte", "eventYN", "treat", "age", "male")

sim_data_nPH <- rbind(sim_data_nPH_t0, sim_data_nPH_t1)
#0.5 and 4 is parameters for unifirm distribution, change it following Table 2.1
to reach different censoring proportion
censoring <- runif(400, 0.5, 4)
  for(i in 1:400){
    if (sim_data_nPH$obs_tte[i] <= censoring[i]){
      sim_data_nPH[i, 3] <- 1
    }
    else {
      sim_data_nPH[i, 3] <- 0
      sim_data_nPH[i, 2] <- censoring[i]
    }

  }
  sim_data_nPH[which(sim_data_nPH$obs_tte > 3), 3] <- 0
  sim_data_nPH[which(sim_data_nPH$obs_tte > 3), 2] <- 3
  return(sim_data_nPH)
}

Generating non-proportional hazards data with exponential censoring:

data_nPH_cen <- function(lamb0, lamb1, alpha0, alpha1){
  covs_t0 <- data.frame(id = 1:200, age = (round(runif(200, 20, 60))-40)/20,
```

```
  gender = rbinom(200, 1, 0.5))

sim_data_nPH_t0 <- simsurv(dist = "weibull", lambdas = lamb0, gammas = alpha0,
x = covs_t0, betas = c(age = -0.5, gender = -0.5))

covs_t1 <- data.frame(id = 201:400, age = (round(runif(200, 20, 60))-40)/20,
gender = rbinom(200, 1, 0.5))

sim_data_nPH_t1 <- simsurv(dist = "weibull", lambdas = lamb1, gammas = alpha1,
x = covs_t1, betas = c(age = -0.5, gender = -0.5))

#combine data
sim_data_nPH_t0 <- cbind(sim_data_nPH_t0, rep(0, 200), covs_t0[, 2:3])
names(sim_data_nPH_t0) <- c("id", "obs_tte", "eventYN", "treat", "age", "male")

sim_data_nPH_t1 <- cbind(sim_data_nPH_t1, rep(1, 200), covs_t1[, 2:3])
names(sim_data_nPH_t1) <- c("id", "obs_tte", "eventYN", "treat", "age", "male")

sim_data_nPH <- rbind(sim_data_nPH_t0, sim_data_nPH_t1)
#0.725 is the rate parameter for exponential distribution, change it following Table 2.1
to reach different censoring proportion
censoring <- rexp(400, 0.725)
  for(i in 1:400){
    if (sim_data_nPH$obs_tte[i] <= censoring[i]){
      sim_data_nPH[i, 3] <- 1
    }
    else {
      sim_data_nPH[i, 3] <- 0
      sim_data_nPH[i, 2] <- censoring[i]
    }

  }
  sim_data_nPH[which(sim_data_nPH$obs_tte > 3), 3] <- 0
  sim_data_nPH[which(sim_data_nPH$obs_tte > 3), 2] <- 3
  return(sim_data_nPH)

}
```

```
Generating proportional hazards data without censoring:

#lamb and alpha is parameters of Weibull distribution for the baseline group
data_PH <- function(lamb, alpha){
  ages <- round(runif(400, 20, 60))
  covs <- data.frame(id = 1:400, treat = rep(c(0,1), 200), age = (ages-40)/20,
  gender = rbinom(400, 1, 0.5))

  sim_data_PH <- simsurv(dist = "weibull", lambdas = lamb, gammas = alpha,
  x = covs, betas = c(treat = -0.5, age = -0.5, gender = -0.5))
  sim_data_PH <- cbind(sim_data_PH, covs[, 2:4])
names(sim_data_PH) <- c("id", "obs_tte", "eventYN", "treat", "age", "male")

return(sim_data_PH)
}

Generating proportional hazards data with uniform censoring:
data_PH_cen <- function(lamb, alpha){
  ages <- round(runif(400, 20, 60))
  covs <- data.frame(id = 1:400, treat = rep(c(0,1), 200), age = (ages-40)/20,
  gender = rbinom(400, 1, 0.5))

  sim_data_PH <- simsurv(dist = "weibull", lambdas = lamb, gammas = alpha,
  x = covs, betas = c(treat = -0.5, age = -0.5, gender = -0.5))

  sim_data_PH <- cbind(sim_data_PH, covs[, 2:4])
names(sim_data_PH) <- c("id", "obs_tte", "eventYN", "treat", "age", "male")
#0.5 and 4 is parameters for unifirm distribution, change it following Table 2.1
to reach different censoring proportion
  censoring <- runif(400, 0.5, 4)
  for(i in 1:400){
    if (sim_data_PH$obs_tte[i] <= censoring[i]){
      sim_data_PH[i, 3] <- 1
    }
    else {
      sim_data_PH[i, 3] <- 0
      sim_data_PH[i, 2] <- censoring[i]
    }
```

```
  }
  sim_data_PH[which(sim_data_PH$obs_tte > 3), 3] <- 0
  sim_data_PH[which(sim_data_PH$obs_tte > 3), 2] <- 3
  return(sim_data_PH)

}
```

Generating proportional hazards data with exponential censoring:

```
data_PH_cen <- function(lamb, alpha){
  ages <- round(runif(400, 20, 60))
  covs <- data.frame(id = 1:400, treat = rep(c(0,1), 200), age = (ages-40)/20,
  gender = rbinom(400, 1, 0.5))

  sim_data_PH <- simsurv(dist = "weibull", lambdas = lamb, gammas = alpha,
  x = covs, betas = c(treat = -0.5, age = -0.5, gender = -0.5))

  sim_data_PH <- cbind(sim_data_PH, covs[, 2:4])
names(sim_data_PH) <- c("id", "obs_tte", "eventYN", "treat", "age", "male")
#0.725 is the rate parameter for exponential distribution, change it following Table 2.1
to reach different censoring proportion
  censoring <- rexp(400, 0.725)
  for(i in 1:400){
    if (sim_data_PH$obs_tte[i] <= censoring[i]){
      sim_data_PH[i, 3] <- 1
    }
    else {
      sim_data_PH[i, 3] <- 0
      sim_data_PH[i, 2] <- censoring[i]
    }

  }
  sim_data_PH[which(sim_data_PH$obs_tte > 3), 3] <- 0
  sim_data_PH[which(sim_data_PH$obs_tte > 3), 2] <- 3
  return(sim_data_PH)

}
```

A.3 code for Kaplan-Meier method

Basic functions

```
#function used to calculate area under stepped function
RMST_KM <- function(KM_treat, t_lim){
  int <- (KM_treat$time[1] - 0)*1 #1 is the initial survival number: 1

  for(i in 2:(length(KM_treat$time))){
    if(KM_treat$time[i] < t_lim){
      int <- int + (KM_treat$time[i] - KM_treat$time[i-1])*KM_treat$surv[i-1]
    }
    else{
      int <- int + (t_lim - KM_treat$time[i-1])*KM_treat$surv[i-1]
      break
    }
  }
  return(int)
}

#function for getting survival results for treatment group
#''lamb0, lamb1, alpha0, alpha1'' and ''data_nPH(lamb0, lamb1, alpha0, alpha1)''
is changable to any data generting function in the data generation subsection

MC_KM1 <- function(lamb0, lamb1, alpha0, alpha1){
  sim_data_nPH <- data_nPH(lamb0, lamb1, alpha0, alpha1)

KM_t1 <- survfit(Surv(obs_tte, eventYN)~treat, data = sim_data_nPH)[2, ]

#construct data frame to save result
KM_result_t1 <- data.frame(time = KM_t1$time, surv = KM_t1$surv, n.event = KM_t1$n.event)

KM_event_t1 <- KM_result_t1[KM_result_t1$n.event != 0, ]

return(KM_event_t1)
}

#function for getting survival results for control group
```

```
MC_KM0 <- function(lamb0, lamb1, alpha0, alpha1){
  sim_data_nPH <- data_nPH(lamb0, lamb1, alpha0, alpha1)

KM_t0 <- survfit(Surv(obs_tte, eventYN)~treat, data = sim_data_nPH)[1, ]

#construct data frame to save result
KM_result_t0 <- data.frame(time = KM_t0$time, surv = KM_t0$surv, n.event = KM_t0$n.event)

KM_event_t0 <- KM_result_t0[KM_result_t0$n.event != 0, ]

return(KM_event_t0)
}

#function to get final mean RMST difference estimation and standard deviation results
over 1000 replication
#t_lim is the selected time limit

result_KM_diff <- function(t_lim){
  r <- NULL
  for(i in 1: 1000){
    r[i] <- RMST_KM(result_MC_KM1[, i], t_lim) - RMST_KM(result_MC_KM0[, i], t_lim)
  }
  RMST_MC_KM <- mean(r)
  sd_KM <- sd(r, na.rm = T)
  return(c(RMST_MC_KM, sd_KM))
}
```

Running results

```
#set.seed(415) is for repeatability
set.seed(415)
result_MC_KM1 <- replicate(1000, MC_KM1(lamb0, lamb1, alpha0, alpha1))
set.seed(415)
result_MC_KM0 <- replicate(1000, MC_KM0(lamb0, lamb1, alpha0, alpha1))

result_KM_diff <- result_KM_diff(t_lim)
```

codes for confidence interval

Here, we offer functions to calculate confidence interval for a single datasets. If the object is to estimate confidence interval for multiple datasets, just replicate the following propcedures.

```
RMST_KM_tr1 <- MC_KM1(lamb0, lamb1, alpha0, alpha1)
RMST_KM_tr0 <- MC_KM0(lamb0, lamb1, alpha0, alpha1)

RMST_SE_KM <- function(KM_treat, t_lim){
  SE <- 0
  for(i in 1:length(which(KM_treat$time <= t_lim))){
    SE <- SE + (RMST_KM(KM_treat, t_lim) - RMST_KM(KM_treat, KM_treat$time[i]))^2*
    (KM_treat$n.event[i]/(KM_treat$atrisk[i]*(KM_treat$atrisk[i] - KM_treat$n.event[i])))
  }
  SE <- sqrt(SE)
  return(SE)
}

SE_t0_KM <- RMST_SE_KM(RMST_KM_tr0, t_lim)

SE_t1_KM <- RMST_SE_KM(RMST_KM_tr1, t_lim)

#0.95 confidence interval for control group
KM_CI_low_tr0 <- RMST_KM(RMST_KM_tr0, t_lim) - qnorm(0.975)*SE_t0_KM

KM_CI_up_tr0 <- RMST_KM(RMST_KM_tr0, t_lim) + qnorm(0.975)*SE_t0_KM

#Confidence interval for treatment group is similar
```

A.4 code for Cox model method

Basic functions

```
#function used to calculate area under stepped function, same as Kaplan-Meier
RMST_KM <- function(KM_treat, t_lim){
  int <- (KM_treat$time[1] - 0)*1 #1 is the initial survival number: 1
```

```
  for(i in 2:(length(KM_treat$time))){
    if(KM_treat$time[i] < t_lim){
      int <- int + (KM_treat$time[i] - KM_treat$time[i-1])*KM_treat$surv[i-1]
    }
    else{
      int <- int + (t_lim - KM_treat$time[i-1])*KM_treat$surv[i-1]
      break
    }
  }
  return(int)
}

#function to collect survival results for treatment group
MC_Cox1 <- function(lamb0, lamb1, alpha0, alpha1){
  sim_data_nPH <- data_nPH(lamb0, lamb1, alpha0, alpha1)
  #formula in Surv()~... here is adjusted for all covariates and stratified by treat.
  Changable for other situations.

  coxfit <- coxph(Surv(obs_tte, eventYN) ~ age + male + strata(treat), ties = ''breslow''
  data = sim_data_nPH)
  #[2,] select strata 2 which is results for treatment group
  Cox_treat1 <- survfit(coxfit, newdata = data.frame(age = 0, male = 0.5))[2, ]

#save result in dataframe
result_Cox_t1 <- data.frame(time = Cox_treat1$time, surv = Cox_treat1$surv,
n.event = Cox_treat1$n.event, n.cersor = Cox_treat1$n.censor,
atrisk = Cox_treat1$n.risk, treat = rep(1, length(Cox_treat1$time)))

return(result_Cox_t1)
}

#function to collect survival results for control group
MC_Cox0 <- function(lamb0, lamb1, alpha0, alpha1){
  sim_data_nPH <- data_nPH(lamb0, lamb1, alpha0, alpha1)
  #formula in Surv()~... here is adjusted for all covariates and stratified by treat.
  Changable for other situations.

  coxfit <- coxph(Surv(obs_tte, eventYN) ~ age + male + strata(treat), ties = ''breslow''
```

```
  data = sim_data_nPH)
  #[1,] select strata 1 which is results for control group
  Cox_treat0 <- survfit(coxfit, newdata = data.frame(age = 0, male = 0.5))[1, ]

#save result in dataframe
result_Cox_t0 <- data.frame(time = Cox_treat0$time, surv = Cox_treat0$surv,
n.event = Cox_treat0$n.event, n.cersor = Cox_treat0$n.censor,
atrisk = Cox_treat0$n.risk, treat = rep(1, length(Cox_treat0$time)))

return(result_Cox_t0)
}

#function for getting final results
result_Cox_diff <- function(t_lim){
  r <- NULL
  for(i in 1: 1000){
    r[i] <- RMST_KM(result_MC_Cox1[, i], t_lim) - RMST_KM(result_MC_Cox0[, i], t_lim)
  }
  RMST_MC_Cox <- mean(r)
  sd_Cox <- sd(r, na.rm = T)
  return(c(RMST_MC_Cox, sd_Cox))
}
```

Running results

```
set.seed(415)
result_MC_Cox1 <- replicate(1000, MC_Cox1(lamb0, lamb1, alpha0, alpha1))
set.seed(415)
result_MC_Cox0 <- replicate(1000, MC_Cox0(lamb0, lamb1, alpha0, alpha1))

Cox_diff <- result_Cox_diff(t_lim)
```

codes for confidence interval

Here, we offer functions to calculate confidence interval for a single datasets. If the object is to estimate confidence interval for multiple datasets, just replicate the following propcedures.

```
#bootstrap function for use in latter bootstrap Confidence interval estimation
boot_COX <- function(survdata, indices){
  #indices allow boot to select sample, no need to insert yourself
  d <- survdata[indices, ]
  #strata(treat) when generalized for non_PH
  coxfit <- coxph(Surv(obs_tte, eventYN) ~ treat, data = d)
  #survival function for treatment group
  Cox_treat0 <- survfit(coxfit, newdata = data.frame(treat=0))
  Cox_treat1 <- survfit(coxfit, newdata = data.frame(treat=1))

  result_Cox_tr0 <- data.frame(time = Cox_treat0$time, surv = Cox_treat0$surv,
  n.event = Cox_treat0$n.event, n.cersor = Cox_treat0$n.censor,
  atrisk = Cox_treat0$n.risk, treat = rep(0, length(Cox_treat0$time)))

  result_Cox_tr1 <- data.frame(time = Cox_treat1$time, surv = Cox_treat1$surv,
  n.event = Cox_treat1$n.event, n.cersor = Cox_treat1$n.censor,
  atrisk = Cox_treat1$n.risk, treat = rep(1, length(Cox_treat1$time)))

  #set a time limit
  t_lim <- 2
  RMST_tr0_Cox <- RMST_KM(result_Cox_tr0, t_lim)
  RMST_tr1_Cox <- RMST_KM(result_Cox_tr1, t_lim)

  return(c(RMST_tr0_Cox, RMST_tr1_Cox))
}

#resampling can be controlled by set.seed
set.seed(415)
#get bootstrap result, R is the number of sampling
#data is a single dataset
boot_result_Cox <- boot(data=data, statistic = boot_COX, R=500)

#bootstrap 0.95 CI, choose the basic bootstrap method (non-studentized),
change 'type' to choose other bootstrap method
#use index to choose control group
CI_t0_Cox <- boot.ci(boot_result_Cox, type = "basic", index = 1)$basic[4:5]
#use index to choose treatment group
CI_t1_Cox <- boot.ci(boot_result_Cox, type = "basic", index = 2)$basic[4:5]
```

```
#the estimated standard error of RMST using bootstrap
#control group
sd(boot_result_Cox$t[, 1])
#treatment group
sd(boot_result_Cox$t[, 2])
```

A.5 code for Flexible parametric method

Basic functions

```
#function for integration calculation
survfun_fp <- function(t, fp, newdata){
  survival <- summary(fp, newdata = newdata, type="survival", t = t)[[1]][, 2]
  return(survival)
}

#function for data fitting
MC_FP <- function(lamb0, lamb1, alpha0, alpha1){
  sim_data_nPH <- data_nPH(lamb0, lamb1, alpha0, alpha1)
  #optimisation in felxsurvsplines may not converged in some sample
  #''treat + gamma1(treat) + gamma2(treat) + gamma3(treat) + gamma4(treat)''
  is the way to adjust for non-PH situation
  #for PH, simple ''treat'' in formula will do the work
  #as the initial values for parameters is not given here. For some sample,
  automatic initial values will cause log(0), error shows up. The try() function
  is convinient when multiple runs are conducted. It will save the error message
  and keep replication continue if one iteration fail

fp_cova <- try({flexsurvspline(Surv(obs_tte, eventYN) ~ treat + gamma1(treat) +
gamma2(treat) + gamma3(treat) + gamma4(treat) + age + male,
data=sim_data_nPH, k = 3, scale = "hazard")}, silent = T)
return(fp_cova)
}

#function to get final result, values in ‘''newdata'' is changable according to
different target population
```

```
result_FP_diff <- function(t_lim){
  newdata1 <- c(1, 0, 0.5)
names(newdata1) <- c("treat", "age", "male")
newdata1 <- t(as.matrix(newdata1))

newdata0 <- c(0, 0, 0.5)
names(newdata0) <- c("treat", "age", "male")
newdata0 <- t(as.matrix(newdata0))
  RMST_fp_cova1 <- NULL
  RMST_fp_cova0 <- NULL
  RMST_fp_diff <- NULL
  for(i in 1:1000){
    RMST_fp_cova1[i] <- try({integrate(survfun_fp, lower=0, upper=t_lim,
    fp = result_MC_FP[[i]], newdata = newdata1)$value}, silent = T)
    RMST_fp_cova0[i] <- try({integrate(survfun_fp, lower=0, upper=t_lim,
    fp = result_MC_FP[[i]], newdata = newdata0)$value}, silent = T)
  }
  RMST_fp_diff <- as.numeric(RMST_fp_cova1) - as.numeric(RMST_fp_cova0)
  RMST_MC_FP <- mean(RMST_fp_diff, na.rm = T)
  sd_FP <- sd(RMST_fp_diff, na.rm = T)
  return(c(RMST_MC_FP, sd_FP))
}
```

Running results

```
set.seed(415)
result_MC_FP <- NULL
for(i in 1:1000){
  result_MC_FP[[i]] <- MC_FP(lamb0, lamb1, alpha0, alpha1)
}

FP_diff <- result_FP_diff(t_lim)
```

codes for confidence interval

```
#function for bootstrap
boot_FP <- function(survdata, indices, newdata, t_lim){
```

```
  d <- survdata[indices, ]
  try({fp_knots <- flexsurvspline(Surv(obs_tte, eventYN) ~ treat + gamma1(treat) +
  gamma2(treat) + gamma3(treat) + gamma4(treat) +age + male,  data=d, k = 3,
  scale = "hazard")}, silent = T)
  try({RMST_fp <- integrate(survfun_fp, lower=0, upper=t_lim, FP_fitting = fp_knots,
  newdata = newdata)}, silent = T)
  return(RMST_fp$value)
}

#here we calculate confidence interval for two treatment groups seperately
data_treat0 <- data[which(data$treat == 0), ]
data_treat1 <- data[which(data$treat == 1), ]

results_FP_t0 <- boot(data=data, statistic = boot_FP, newdata = data_treat0[1, ],
t_lim = t_lim, R=1000)
results_FP_t1 <- boot(data=data, statistic = boot_FP, newdata = data_treat1[1, ],
t_lim = t_lim, R=1000)

#omit the NA values in boot_FP result before using boot.ci
results_FP_t0$t <- na.omit(results_FP_t0$t)
results_FP_t1$t <- na.omit(results_FP_t1$t)

#bootstrap 0.95 CI
#control group
CI_t0_FP <- boot.ci(results_FP_t0, type = "basic")$basic[4:5]

#treatment group
CI_t1_FP <- boot.ci(results_FP_t1, type = "basic")$basic[4:5]

#the estimated standard error
#control group
sd(results_FP_t0$t)

#treatment group
sd(results_FP_t1$t)
```

A.6 code for POGEE

Basic functions

```
#POGEE follow a different logic, the parameter setting is a little bit different.
diff_POGEE <- function(t_lim, lamb0, lamb1, alpha0, alpha1){

sim_data_nPH <- data_nPH(lamb0, lamb1, alpha0, alpha1)

po_RMST <- pseudomean(time = sim_data_nPH$obs_tte, event = sim_data_nPH$eventYN,
tmax = t_lim)

data_po <- cbind(sim_data_nPH, po = po_RMST)

#fit regression model, can get results adjusted for more covariates
fit_po <- geeglm(po_RMST ~ treat + age + male, data = data_po, id=id,
family = "gaussian", corstr = "independence", scale.fix = F)

diff_POGEE <- as.numeric(predict(fit_po, newdata = data.frame(treat=1,
age=0, male=0.5))) - as.numeric(predict(fit_po, newdata =
data.frame(treat=0, age=0, male=0.5)))

return(diff_POGEE)
}

result_diff_POGEE <- function(t_lim, lamb0, lamb1, alpha0, alpha1){
  result_diff_POGEE <- replicate(1000, diff_POGEE(t_lim, lamb0, lamb1, alpha0, alpha1))
  RMST_diff_POGEE <- mean(result_diff_POGEE)
  diff_sd <- sd(result_diff_POGEE, na.rm = T)
  return(c(RMST_diff_POGEE, diff_sd))
}
```

Running results

```
set.seed(415)
POGEE_diff <- result_diff_POGEE(t_lim, lamb0, lamb1, alpha0, alpha1)
```

codes for confidence interval

```
#bootstrap function for treatment group
boot_PO_tr1 <- function(data, indices){
  d <- data[indices, ]
  #set a time limit as t_lim
  t_lim <- 2
  po_RMST <- pseudomean(time = d$obs_tte, event = d$eventYN, tmax = t_lim)
  data_po <- cbind(d, po = po_RMST)

  #fit regression model, can get results adjusted for more covariates
  #no PH assumption
  fit_po <- geeglm(po_RMST ~ treat + age + male, data = data_po, id=id,
  family = "gaussian", corstr = "independence", scale.fix = F)

  #newdata changable
  POGEE_tr1<-as.numeric(predict(fit_po, newdata = data.frame(treat=1, age=0, male=0.5)))
  return(POGEE_tr1)
}

#bootstrap function for control group
boot_PO_tr0 <- function(data, indices){
  d <- data[indices, ]
  #set a time limit as t_lim
  t_lim <- 2
  po_RMST <- pseudomean(time = d$obs_tte, event = d$eventYN, tmax = t_lim)
  data_po <- cbind(d, po = po_RMST)

  #fit regression model, can get results adjusted for more covariates
  #no PH assumption
  fit_po <- geeglm(po_RMST ~ treat + age + male, data = data_po, id=id,
  family = "gaussian", corstr = "independence", scale.fix = F)

  #newdata changable
  POGEE_tr0<-as.numeric(predict(fit_po, newdata = data.frame(treat=0, age=0, male=0.5)))
  return(POGEE_tr0)
}
```

```
#bootstrap result
results_PO_t0 <- boot(data = data, statistic = boot_PO_tr0, R=1000)
results_PO_t1 <- boot(data = data, statistic = boot_PO_tr1, R=1000)

#bootstrap 0.95 CI
CI_t0_PO <- boot.ci(results_PO_t0, type = "basic")$basic[4:5]
CI_t1_PO <- boot.ci(results_PO_t1, type = "basic")$basic[4:5]

#the estimated standard error
#treatment0
sd(results_PO_t0$t)

#treatment1
sd(results_PO_t1$t)
```

B Appendix : Supplementary plots

B.1 standard deviation plots corresponding to RMST plots in Chapter 3

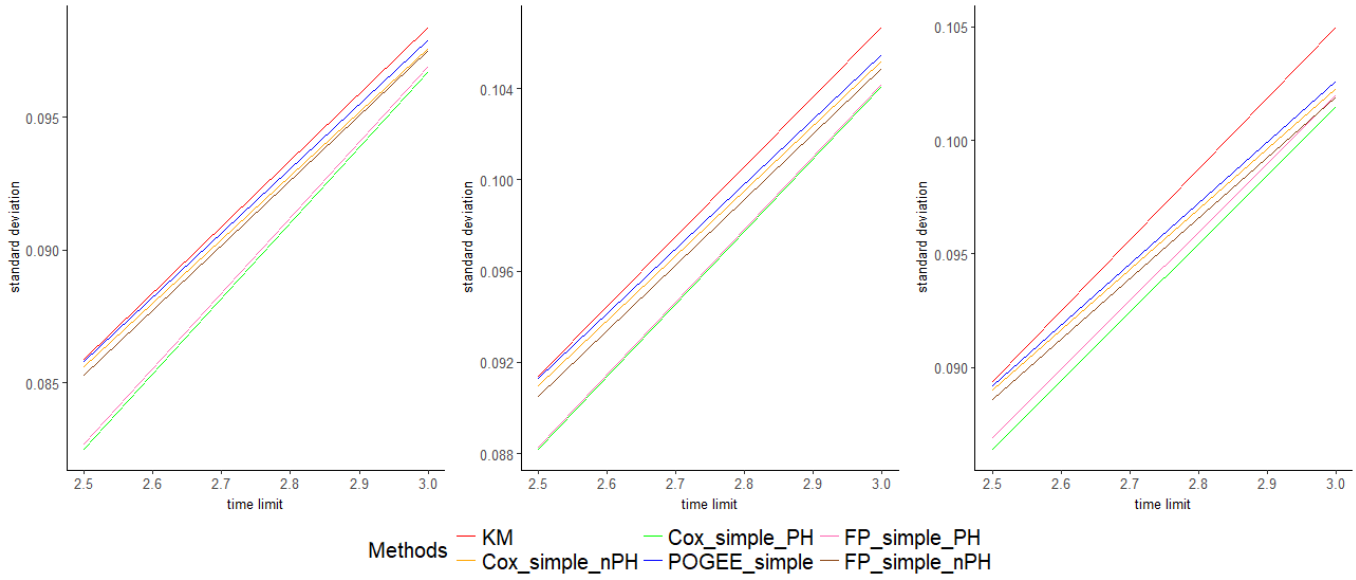


Figure 1: Standard deviation corresponding to Figure 3.3, time range [2.5, 3]

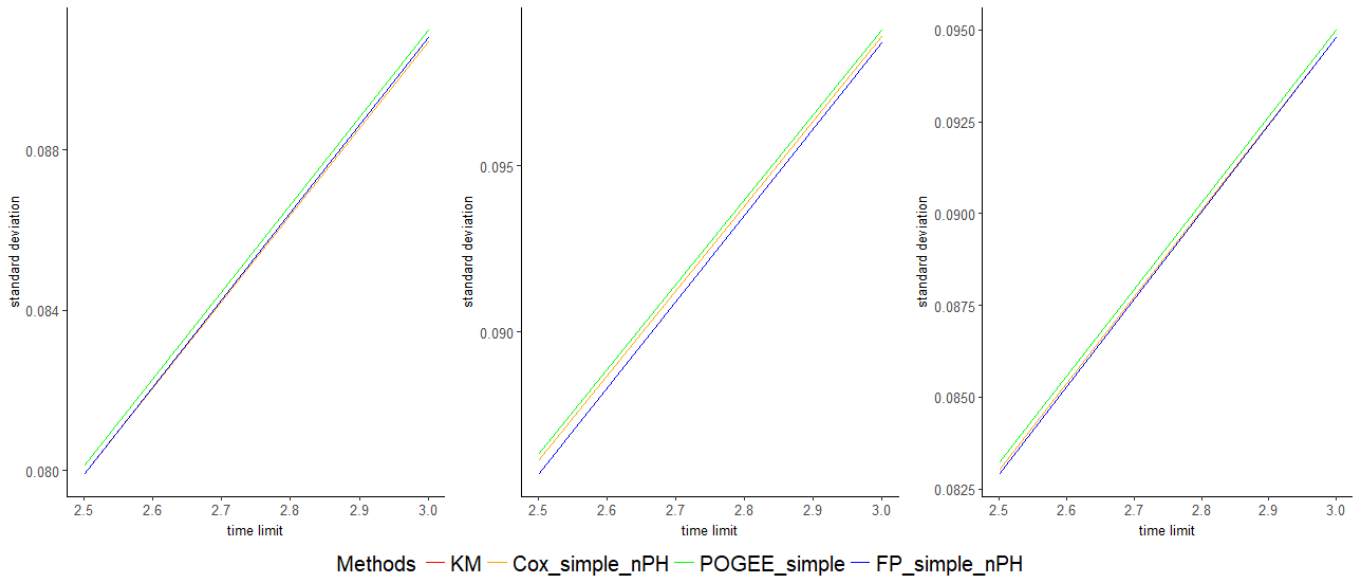


Figure 2: Standard deviation corresponding to Figure 3.4, time range [2.5, 3]

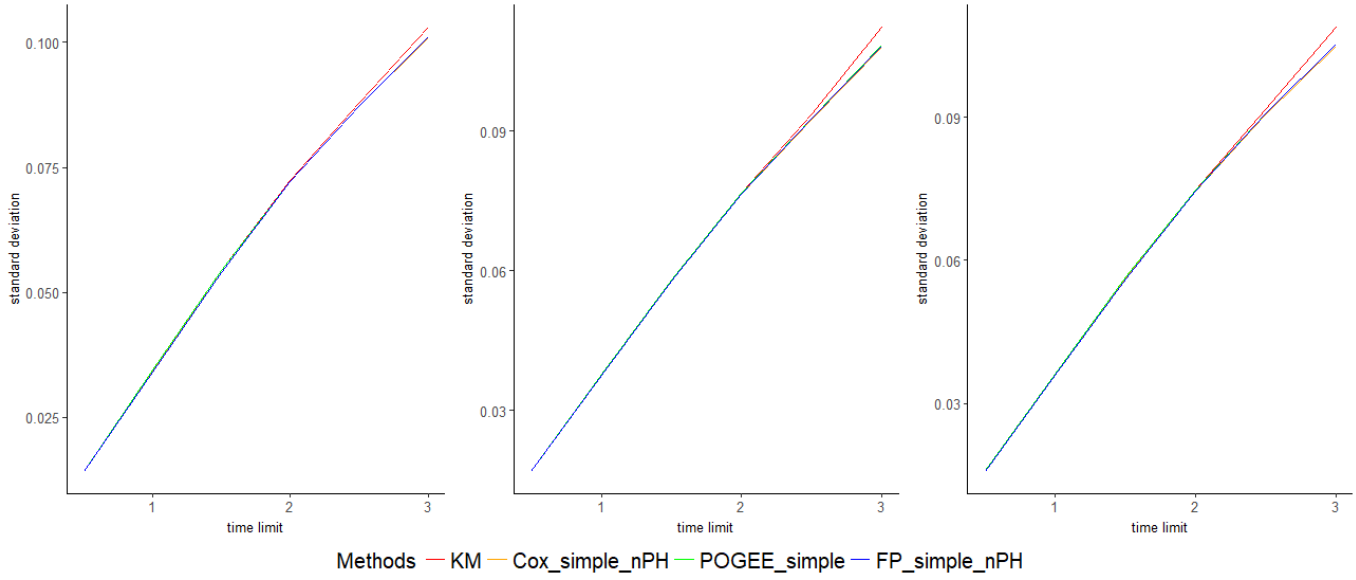


Figure 3: Standard deviation corresponding to Figure 3.8

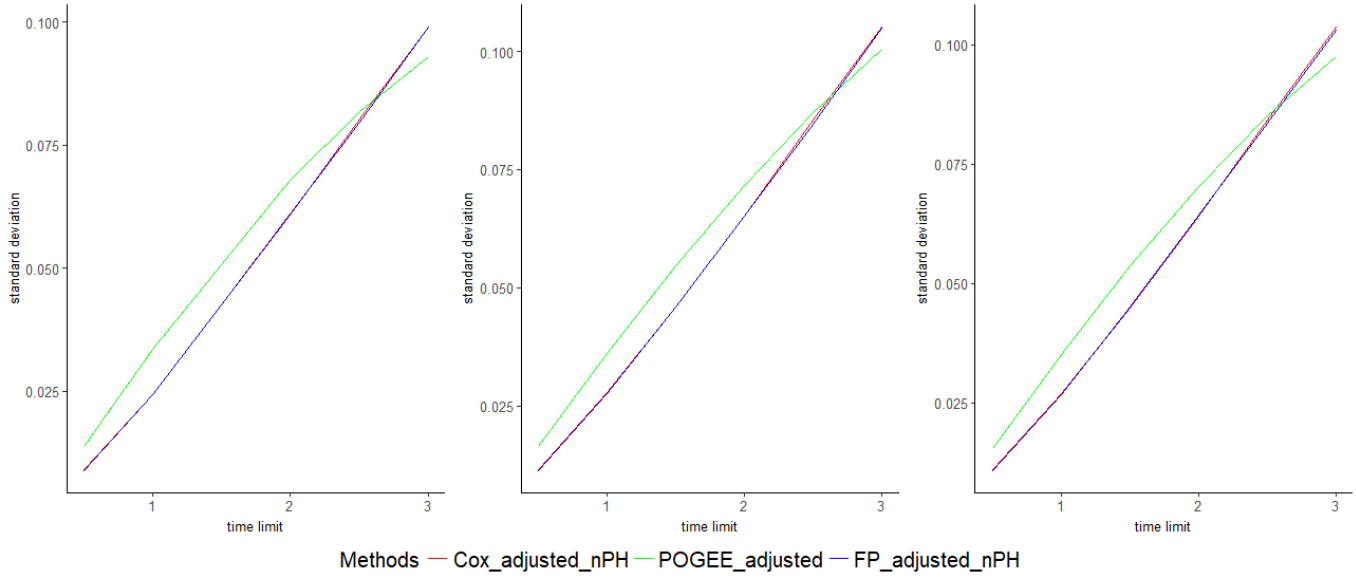


Figure 4: Standard deviation corresponding to Figure 3.9

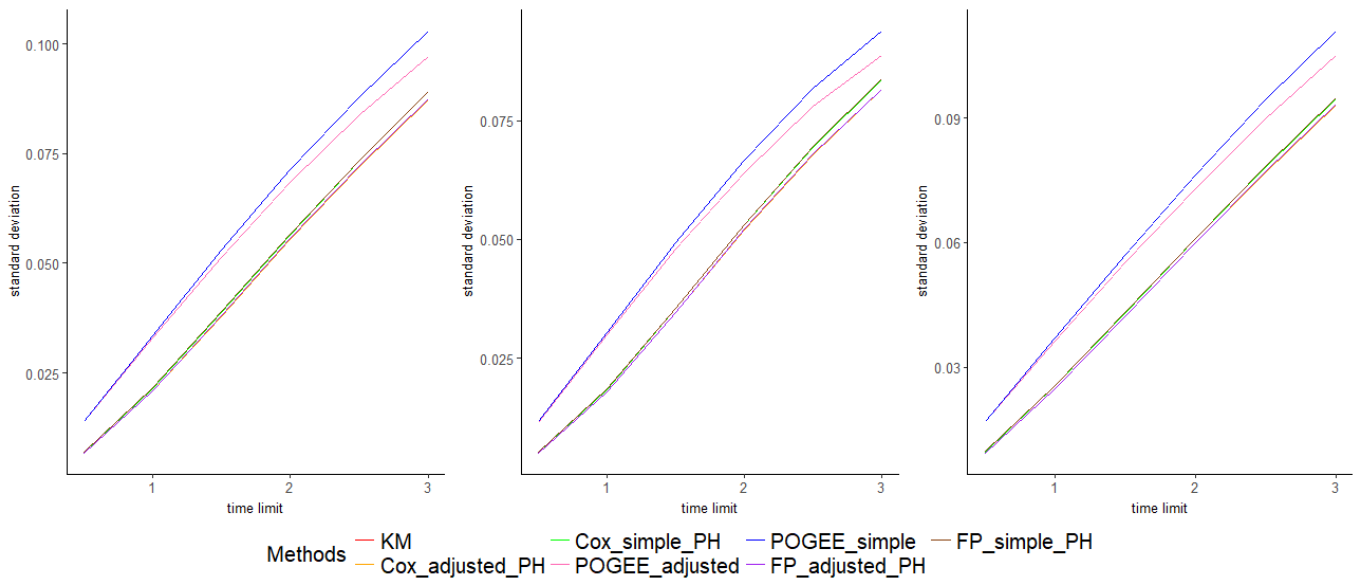


Figure 5: Standard deviation corresponding to Figure 3.10

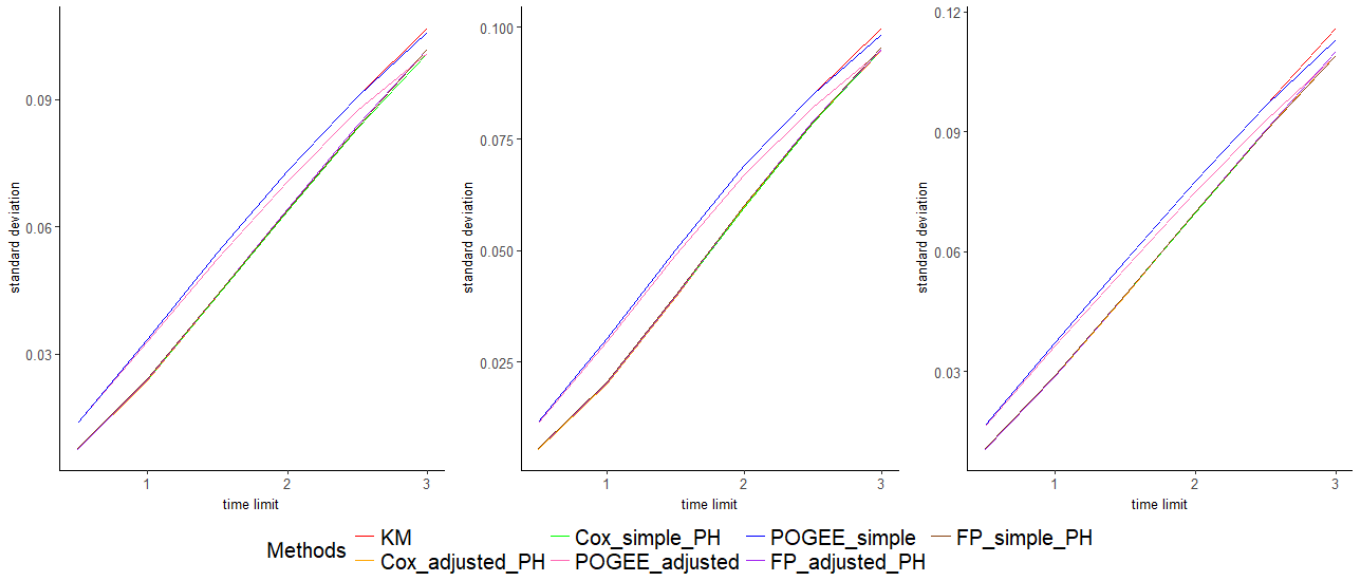


Figure 6: Standard deviation corresponding to Figure 3.14

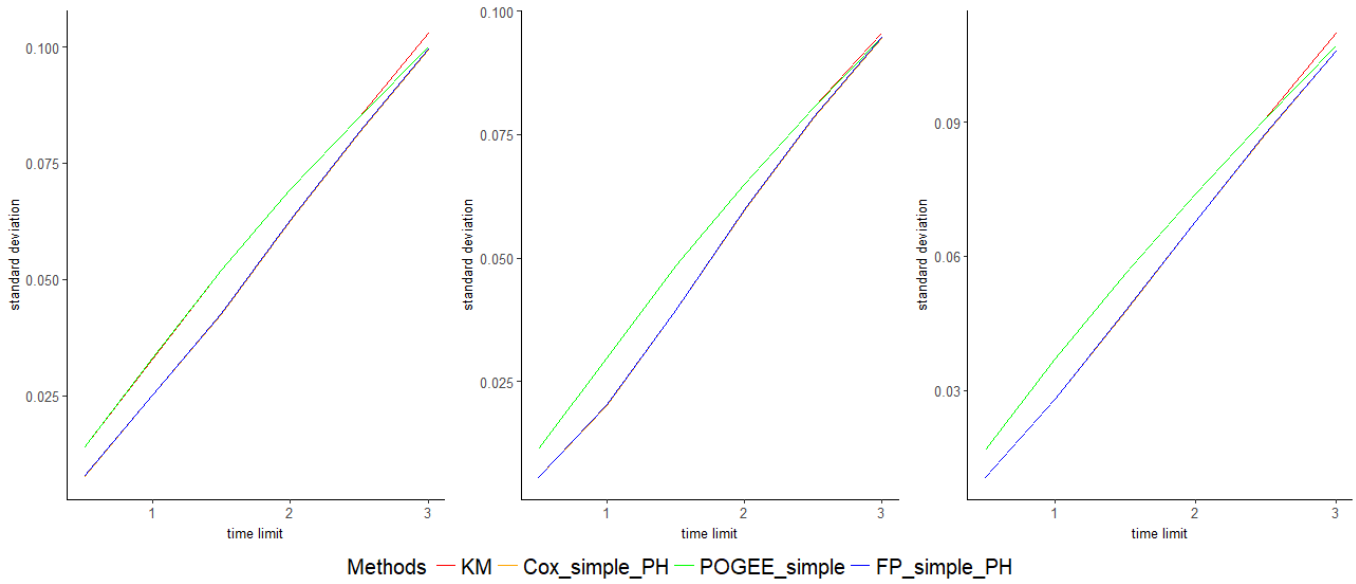


Figure 7: Standard deviation corresponding to Figure 3.13

B.2 extra plots as supplement

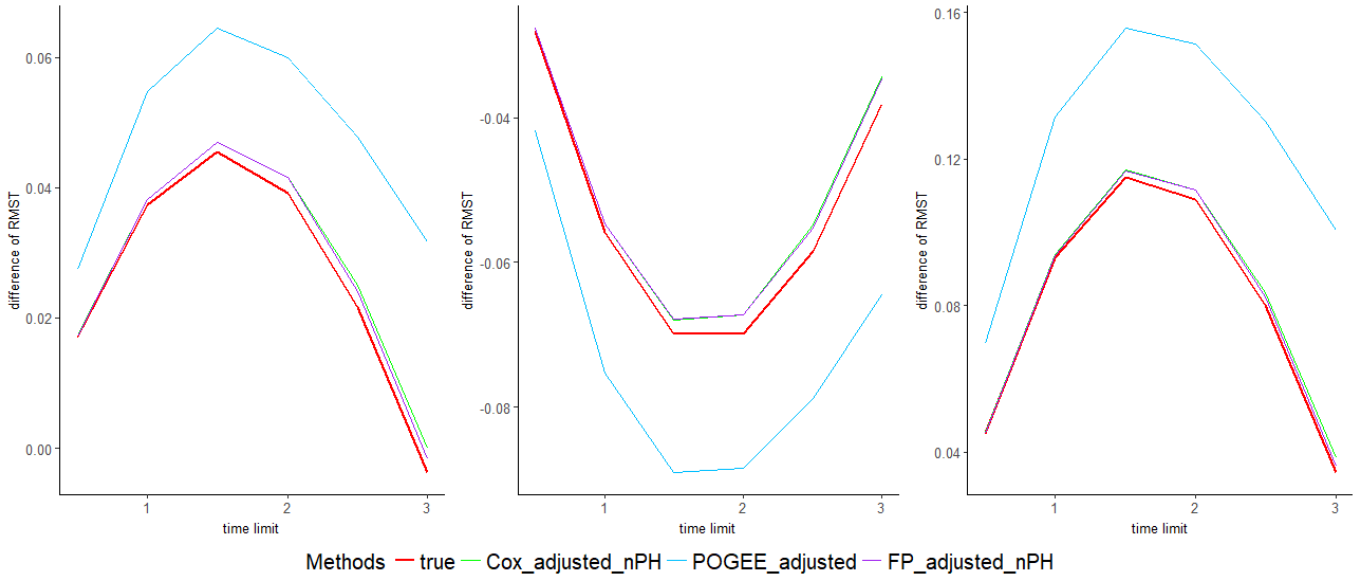


Figure 8: Performance of methods estimating RMST difference for 50 years old male under nPH datasets without censoring. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

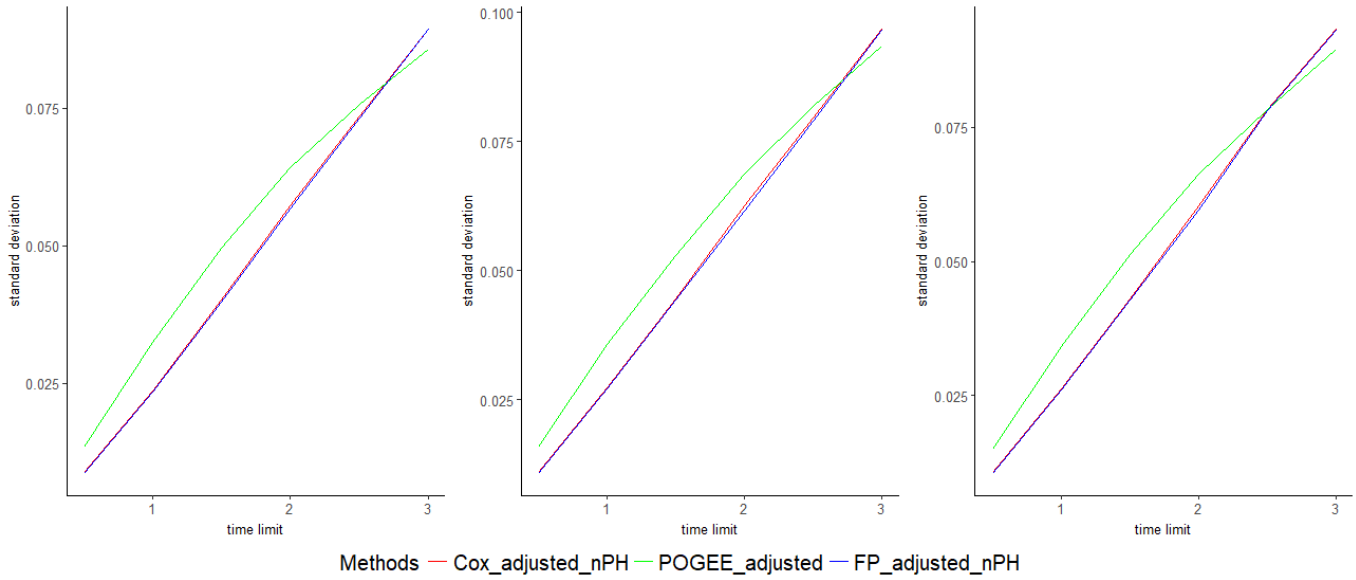


Figure 9: Standard deviation corresponding to above plot

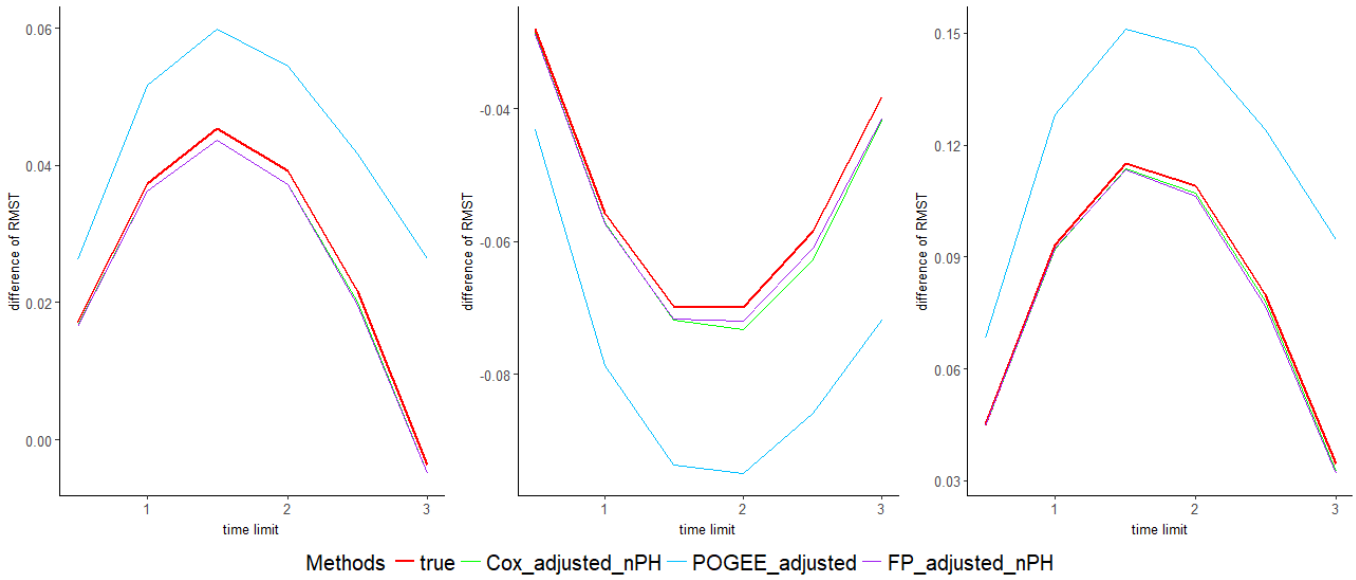


Figure 10: Performance of methods estimating RMST difference for 50 years old male under nPH datasets with overall 50% exponential censoring. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

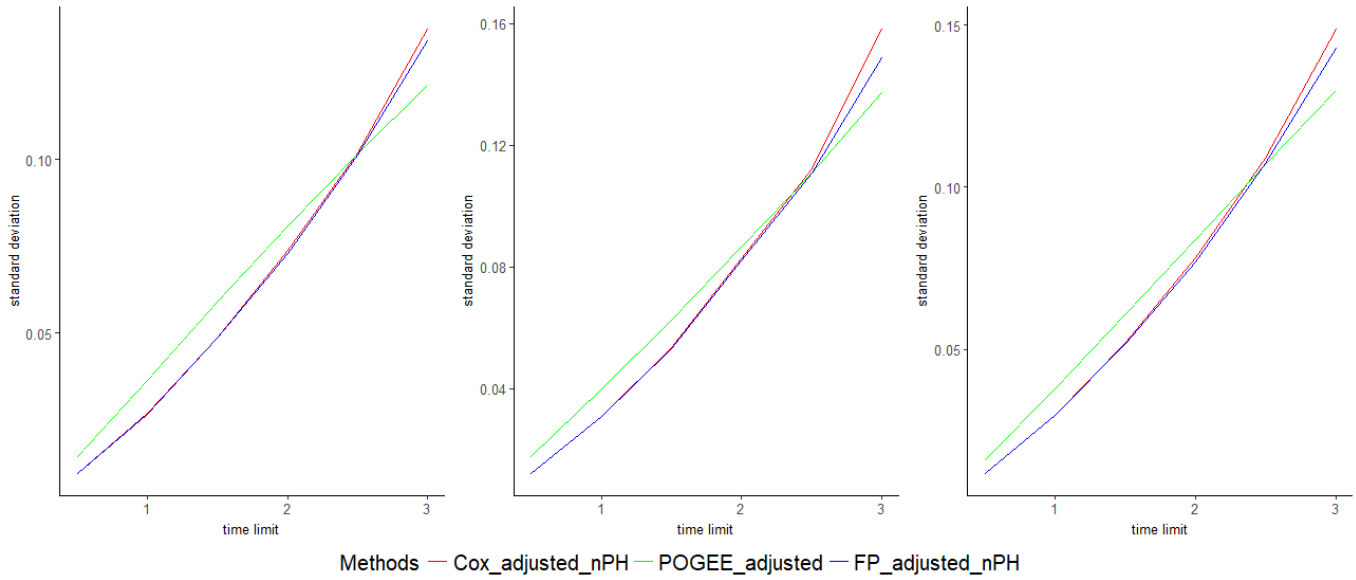


Figure 11: Standard deviation corresponding to above plot

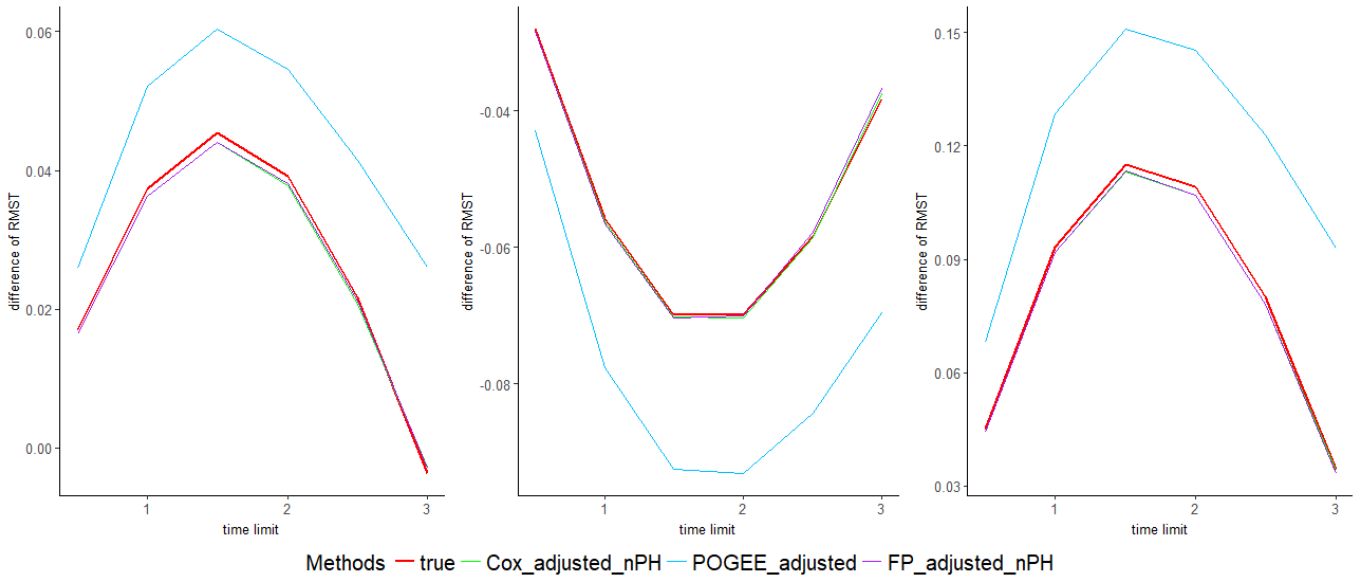


Figure 12: Performance of methods estimating RMST difference for 50 years old male under nPH datasets with overall 25% uniform censoring. Left panel: constant baseline hazard for control group, increasing baseline hazard for treatment group; Middle panel: constant baseline hazard for control group, decreasing baseline hazard for treatment group; Right panel: decreasing baseline hazard for control group, increasing baseline hazard for treatment group

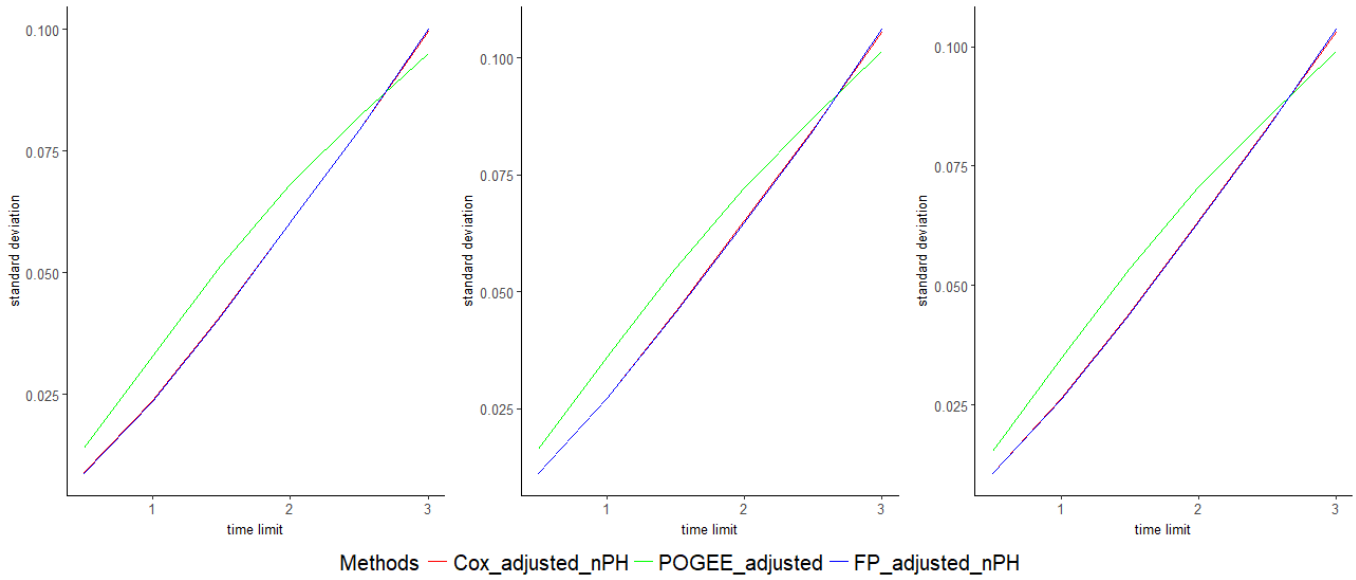


Figure 13: Standard deviation corresponding to above plot

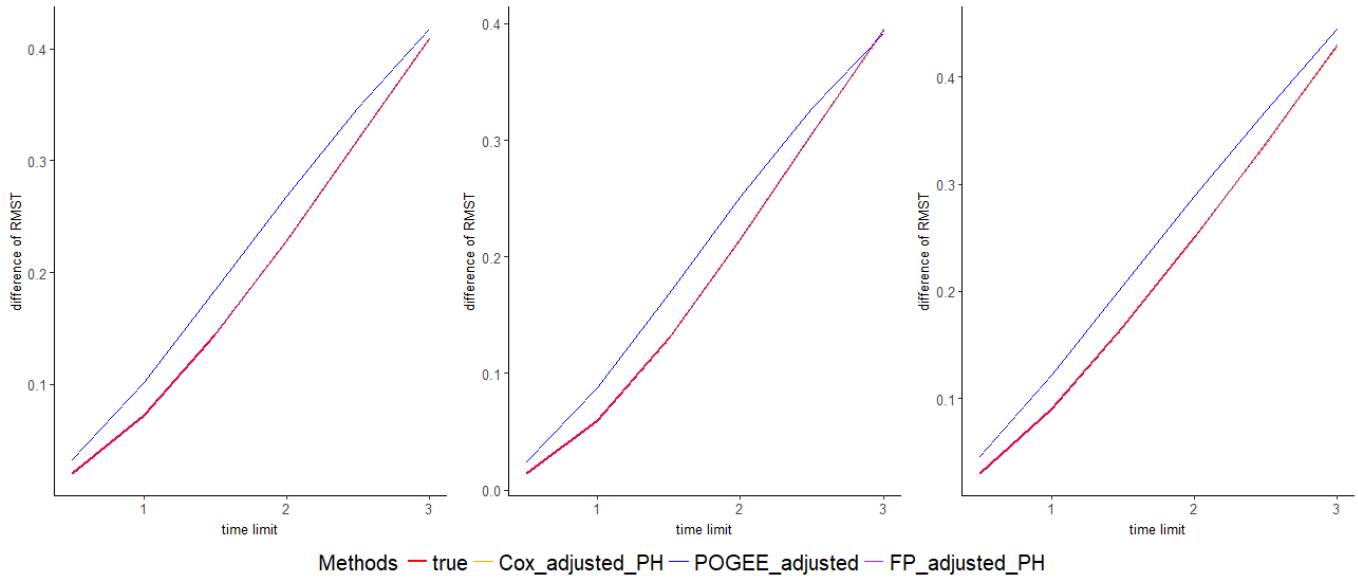


Figure 14: Performance of methods estimating RMST difference for 50 years old male under PH datasets without censoring. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

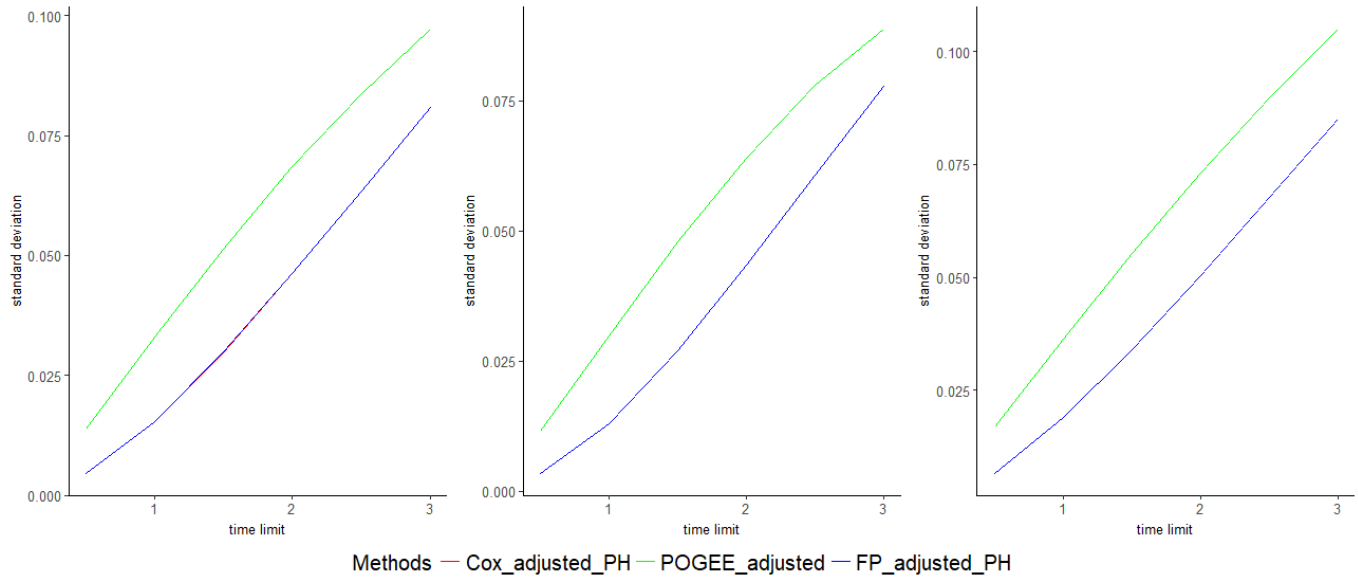


Figure 15: Standard deviation corresponding to above plot

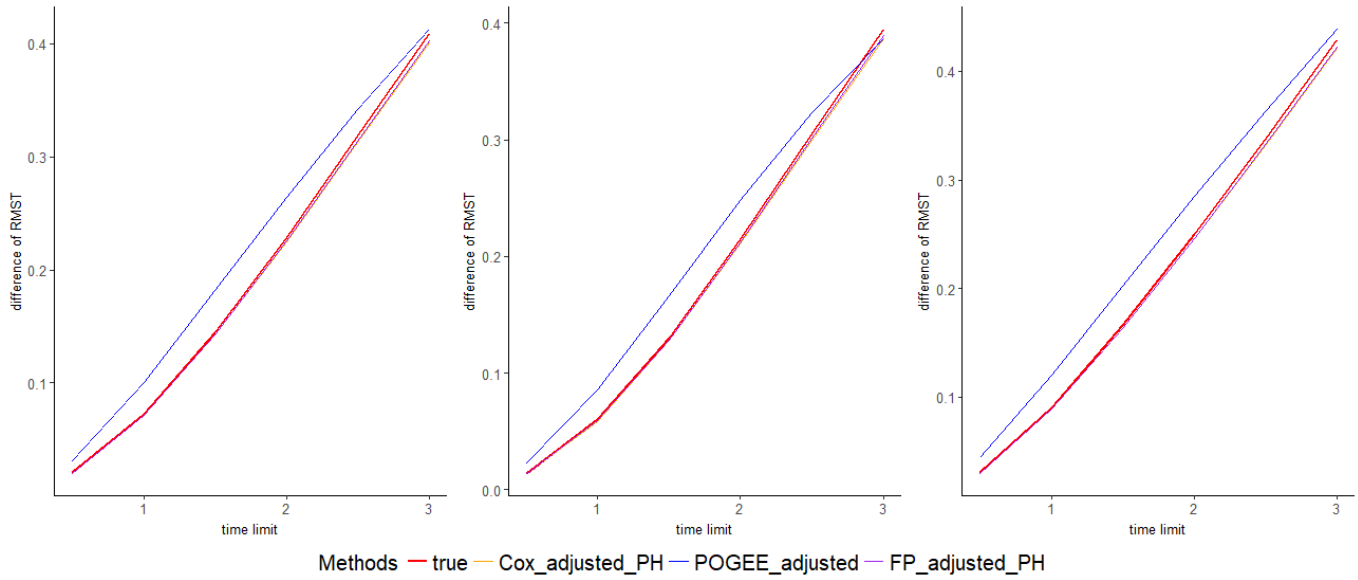


Figure 16: Performance of methods estimating RMST difference for 50 years old male under PH datasets with overall 50% exponential censoring. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

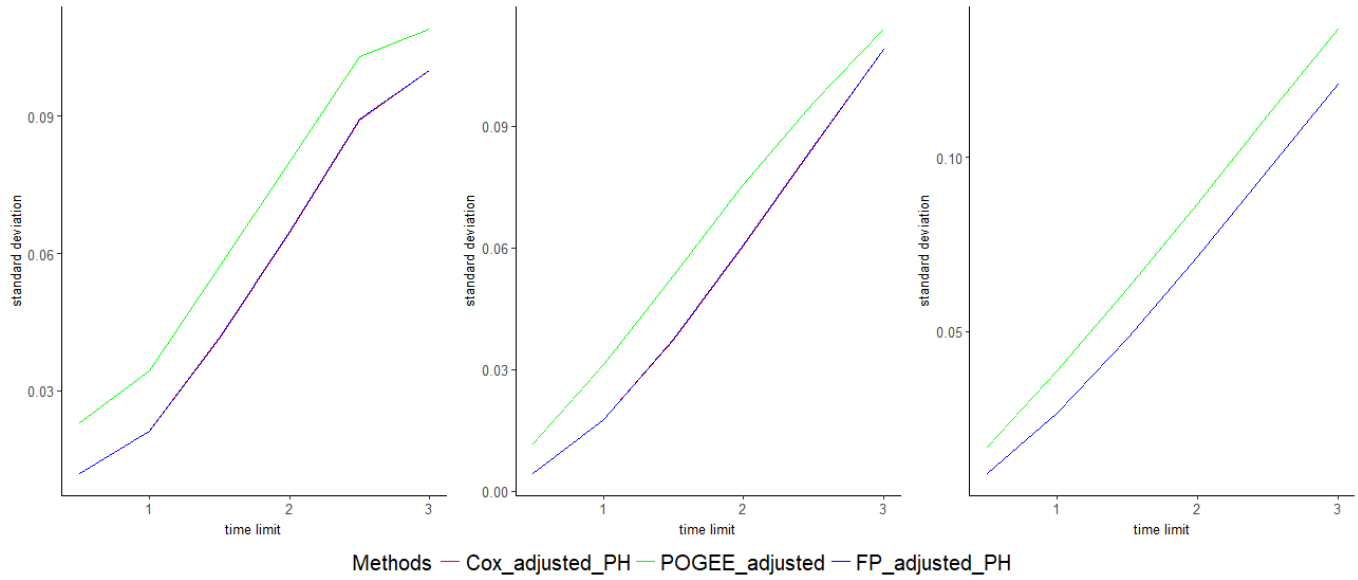


Figure 17: Standard deviation corresponding to above plot

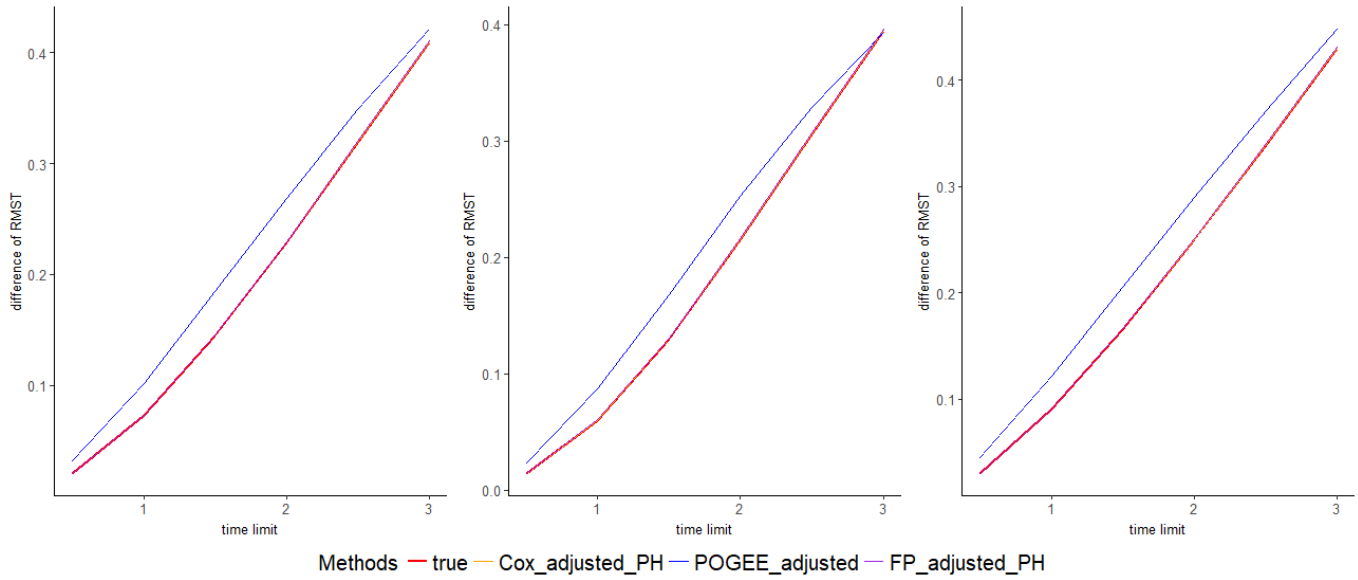


Figure 18: Performance of methods estimating RMST difference for 50 years old male under PH datasets with overall 25% uniform censoring. Left panel: constant baseline hazard; Middle panel: increasing baseline hazard; Right panel: decreasing baseline hazard

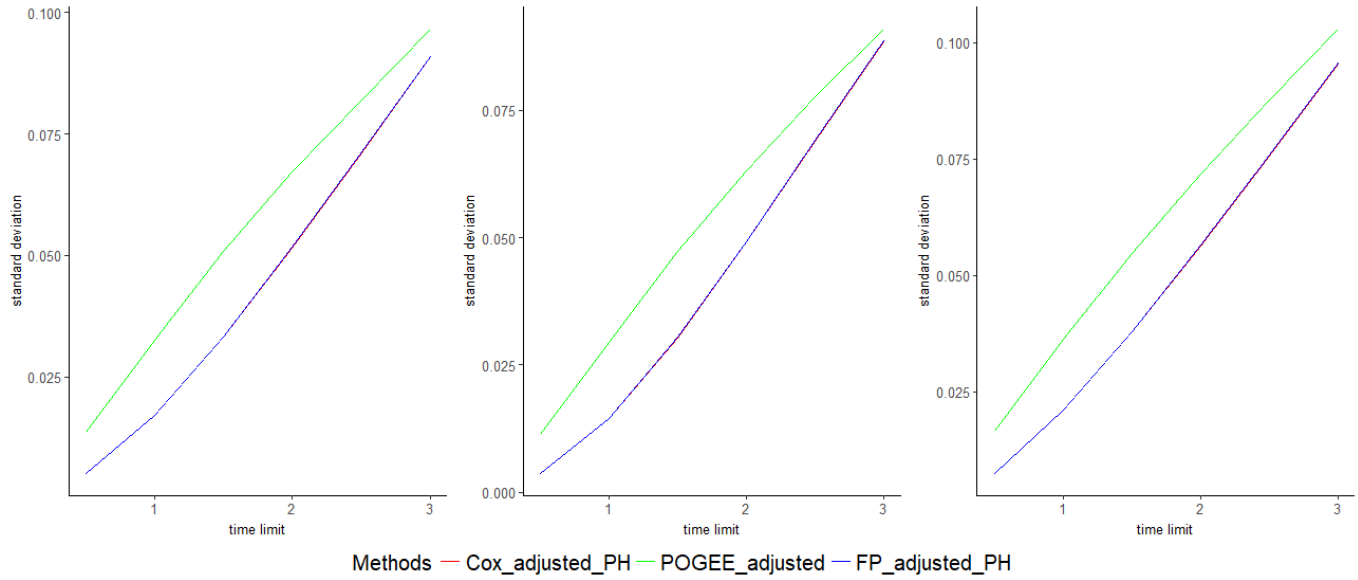


Figure 19: Standard deviation corresponding to above plot