Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Testing the treatment effect in randomized clinical trials with possible non-proportional hazards

Promotor :
Prof. dr. Roel BRAEKERS

Promotor :
Dr. ANDREA CALLEGARO

Belay Belete Anjullo

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Testing the treatment effect in randomized clinical trials with possible non-proportional hazards

Promotor :
Prof. dr. Roel BRAEKERS

Promotor :
Dr. ANDREA CALLEGARO

Belay Belete Anjullo

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

DECLARATION

This thesis is written by me and of my original research work. Wherever contributions of others are found reference to the literature is made.

Signature.......................... Date.............

Belay Belete Student

In our capacity as supervisors of Belay Belete's thesis , we endorse that the above statements are true to the best of our knowledge and it has not already been accepted for any degree, and is also not being concurrently submitted for any other degree and thus permitted to be defended for further evaluation.

Signature...................................... Date..............

Prof. dr. Roel BRAEKERS Internal Supervisor

Signature...................................... Date..............

Dr. Andrea Callegaro External Supervisor

ACKNOWLEDGEMENTS

First of all, I would like to thank almighty God for his uncountable help in my entire life. The completion of this thesis could not have been possible to be appearing in its current form without the support and guidance of several people and organizations. Therefore, my heartfelt gratitude and appreciation goes to my internal supervisor Prof. dr. Roel BRAEKERS and external supervisor dr. Andrea Callegaro for their comments, inspiration and guidance throughout the process of completion of this research work.

I am grateful to the Flemish Interuniversity Council (Vlaamse Interuniversitare Raad (VLIR)) university cooperation development program for sponsoring me financially and Hasselt University graduate program for the opportunity provided for me to pursue masters degree in Biostatistics. I would also like to express my sincere thanks to all my lecturers whose knowledge helped me immensely in the realization of this work. I am indebted to the GSK Vaccines for the opportunity given to me to conduct my thesis within their company.

I would like to thank my friends Andre Honba, Yimer Wasihun, Wondiber Nega, Admasu Geneti and Yonas Ayalew, and all my classmates for helping me get through the difficult times, and for all the emotional support, entertainment and caring they provided. I am grateful to the secretaries especially Ms. Martine Machiels for assisting me in many different ways. I would also like to thank my lovely Merry Biru for her unconditional love and encouragement during my studies.

I would like to extend my deepest thanks to my parents for their lifelong love and support in my entire life time. Lastly, it is my pleasure to thank those who are not mentioned by names but whose support made this thesis a success.

ABSTRACT

Many randomized clinical trials includes right censored time to event data, comparing an experimental treatment with a standard treatment or placebo control. In this comparison, one tests whether the two treatments have the same survival function or equivalently the same hazard function over a given time period in order to evaluate effect of treatment. The methodological development of survival analysis for randomized clinical trials with right-censored data that have had the most profound impact are the log-rank test for comparing the equality of two or more survival distributions, and the Cox proportional hazards model for examining the covariate(s) effects on the hazard function. However, when comparing treatments in terms of their time to event distribution, there may be reason to believe that the hazard curves will cross, and in such cases standard comparison techniques could lead to misleading results (Logan et al, 2008). Hence, in this study, the performance of new methods proposed by Callegaro et al (2014) for testing treatment effect on randomized clinical trials when the proportional hazards assumption is in doubt was evaluated based on simulation studies and on two real datasets. New proposed methods are based on combination of early/late treatment effects obtained from stopped/left truncated Cox or equivalently from extended Cox *and the overall treatment effect from Cox proportional hazards model. These methods were compared with Cox (1972) proportional hazards model, pseudo values regression approach based on mean restricted survival time proposed by Andersen et al (2004) and Klein et al (2007) and extended Cox for the time dependent treatment effect proposed by Putter et al (2005). Type I error rate and power of the proposed tests were illustrated based on simulated data under five possible treatment effect. The results of simulations and real data examples on cancer clinical trials showed that the new proposed methods performed reasonably well in case of crossing survival curves compared to Cox proportional hazards model and pseudo values regression approach based on restricted mean survival time. However, they performed about the same compared to extended Cox model. Furthermore, they performed about the same compared to Cox proportional hazards model and extended Cox under the late treatment effect. Using the proposed methods under proportional hazards alternative did not generally yield dramatic decrease in power compared to the Cox model and they allow to adjust for covariate(s).*

KEY WORDS: *Simulation, Stopped Cox, Kaplan-Meier method, Cox proportional hazards, pseudo values regression approach, extended Cox model*

LIST OF TABLES AND FIGURES

Tables Page

Figures page

1. INTRODUCTION

1.1. Background

Survival analysis has become one of the most widely used statistical tools for analyzing clinical research data. It is specifically concerned with time to event data and is of particular value because of its intrinsic ability to handle censored observations. In the literature, many randomized clinical trials includes right censored time to event data, comparing an experimental treatment with a standard treatment or placebo control in order to evaluate treatment effect (Bain and Engelhardt (1991), Klein and Moeschberger (1997)). In this comparison, one tests whether the two treatments have the same survival function or equivalently the same hazard function over a given follow up time (Zhang and Klein, 1998). The log-rank test is commonly used test statistic for the comparison. Often in these trials, characteristics of the patient and of the tumours that are known before treatment are also recorded. Hence, to study the effect of treatment, Cox proportional hazards model is the most popular choice with advantages of adjusting for baseline and prognostic covariate(s) (Logan *et al*, 2008).

One of the assumptions underlying the Cox model is the assumption of proportional hazards, meaning that the ratio of the hazards for treatment versus control is constant over time (Cox, 1972). Then, the hazards ratio can be expressed as a single number; the hazards ratio of treatment over control. Although not as implicitly assumed as in the Cox regression model, the validity of the log-rank test is also sensitive to the assumption that the hazard ratios for treatment versus control do not change appreciably over time (Putter *et al*, 2005). When studying survival data over a short period of time, the proportional hazards assumption is often a reasonable one. However, in cancer clinical trials with long-term follow-up, it often happens that the hazard ratio changes over time. In the beginning of the study for instance, the experimental treatment may yield better survival, but this effect may be reversed after some time or vice versa (Logan *et al*, 2008). In such a case, the log-rank test for the difference in survival ratios between the treatments will most likely not be significant, because of the contrasting early and late effects of the treatments. If the proportional hazards assumption fails to hold for the treatment or for one or more of the covariates, the results of a Cox model will be misleading. In addition, it is not easy to interpret the hazards ratio resulting from the Cox proportional hazards model because it is a weighted average hazards ratio over the observed follow-up time (Schemper *et al*, 2009, Royston and Parmar, 2011 and Oquigley and Pessione, 1991).

In the literature, to deal with the issues of non-proportional hazards, Putter *et al* (2005) proposed to model the Cox regression model with time-dependent treatment effects. Klein *et al* (2007) proposed to compare survival curves at one fixed time point. Royston and Parmar (2011) proposed to compare restricted mean survival time at a pre-specified fixed time point. Chen and Tsiatis (2001) studied methods for comparing covariate-adjusted restricted mean survival times between two treatment groups. Yang *et al* (2007) proposed testing treatment effect by combining weighted log-rank tests and using empirical likelihood. Logan *et al* (2008) proposed to test two subhypothesis: the hypothesis of equality of Kaplan-Meier survival difference at a pre-specified time point (t_0) and the hypothesis of no difference in the hazards after t_0. Callegaro *et al* (2014) proposed testing treatment effect based on the combination of early (late) and overall treatment effects.

1.2. Statement of the Problem

As many studies indicated, the Cox proportional hazards model is the standard for evaluation of treatment effects on clinical trial data, but when hazards are not proportional, the Cox may not be powerful. Consequently, different approaches have been proposed as alternative to the Cox model in the case of non-proportional hazards. Therefore, this study has attempted to answer the following scientific questions:

a) What are the alternative methods to test the effect of treatment in randomized clinical trial when proportional hazards assumption is in doubt?
b) How is the performance of new methods proposed by Callegaro *et al* (2014) compared to Cox proportional hazards model, pseudo values regression approach based on restricted mean survival time and extended Cox in the situation where proportional hazards assumption is in doubt?

1.3. Objectives of the Study

A comprehensive review of the existing methods for dealing with the problem of non-proportional hazards is provided. It is stressed in the literature that the log-rank or Cox test has optimal power to detect differences in the hazard rates, when the hazard rates are proportional (Klein, Moeschberger, 1997). When these tests are applied to samples from populations where the hazard rate cross, they lack power. Therefore, the main objective of this thesis was to evaluate the performance of newly proposed methods by Callegaro *et al*

(2014) (i.e. methods based on the combination of treatment effects) compared to tests obtained from Cox proportional hazards model, pseudo values regression approach based on restricted mean survival time proposed by Royston and Parmar (2011) and Andersen *et al* (2004) and extended Cox model proposed by Putter *et al* (2005) in order to test treatment effect in randomized clinical trials with possible non-proportional hazards with and without including covariate(s) in the models. This was studied by simulations and two popular real datasets from randomized cancer clinical trials.

1.4. Significance of the Study

This study evaluates the performance of newly proposed methods and offers a breakthrough in the new methods of testing treatment effects in the situations where proportional hazards is in doubt. Therefore, this will increase the bank of knowledge in the field of survival analysis.

2. METHODOLOGY

2.1. Description of the Data

2.1.1. Dataset on gastric cancer trial

In addition to simulations, to illustrate efficiency of newly proposed methods two popular real datasets were considered. Both datasets are taken from the R package survival. The first dataset was on gastric cancer (Stablein, Carter, and Novak 1981) which comes from a controlled clinical trial in patients with advanced non-resectable gastric carcinoma. It was analyzed by MacKenzie and Ha (2007) and Klein and Moeschberger (1997) to exemplify crossing hazards scenario. In this dataset there are two treatment arms: chemotherapy plus radiation and chemotherapy without radiation. There are a total of 90 patients involved in the study and 79 of them are observed events resulting to 12% censoring. The outcome of interest was overall survival time and the objective of the trial was to test if chemotherapy plus radiation is better than chemotherapy without radiation. This dataset was used to exemplify crossing survival curves.

2.1.2. Dataset on bladder cancer trial

The second dataset that was considered to illustrate new proposed methods was coming from a study by Byar (1984) and included patients with superficial bladder tumors removed by transurethral resection. Many patients had multiple tumor recurrences (up to a maximum of 9) during the study, and new tumors were removed at each visit. However, in this study data from 85 individuals in the placebo and thiotepa treatment groups with only the first recurrence was considered and 45% of them are censored. The covariates that were considered are the initial number of tumours and the size (cm) of largest initial tumour. This dataset was used to exemplify the late treatment effect.

2.1.3. Simulation Design

A simulation study was designed to compare the performance of the new proposed methods in terms of their type I error rate and power. Callegaro *et al* (2014) conducted a simulation study to examine the statistical power of their proposed test statistics under a variety of possible situations. They claimed that their proposed test statistics can be used in testing treatment effect, whether or not the underlying proportional hazards assumption was met. Therefore, in this study, a similar simulation setting was carried out to evaluate the power of their proposed test statistics under different possible scenarios and they were compared with some of the existing methods such as pseudo values regression approach based on restricted mean survival time and extended Cox model for time dependent treatment effect. In the simulation design, survival times for treatment groups were generated independently for samples of size 200 subjects per treatment group with 30% of administrative censoring (censoring due to termination of study) using true survival functions presented in Figure 1. This was done under five different scenarios such as: in scenario 1) survival curves are assumed to be identical (i.e., no treatment effect under the null hypothesis), 2) survival curves are assumed to have proportional hazards, 3) survival curves are assumed identical at the beginning, then separate as time goes on (late treatment effect), 4) the two survival curves are separate at beginning, but identical as time goes on (early treatment effect) leading to crossing hazards, and 5) survival curves are assumed to cross. In all scenarios survival times are simulated conditioning on the binary covariate which was generated from Bernoulli distribution considering the follow up period of five years and independent of the censoring times. For each scenario, the data are replicated 1000 times which is the most common choices (Burton *et al,* 2006). The type I error rate and empirical power of the tests are calculated as the proportion of 1000 repeated random samples in which the null hypothesis is rejected at the nominal alpha of 5% with one-sided test statistics under identical survival curves and four different alternative scenarios, respectively with and without including the covariate in the models. Simulations and analyses were done using R software version of R3.1.0.

2.2. Method of Statistical Analysis

2.2.1. Testing for the Treatment Effect based on Pseudo-Values Regression Approach

In survival analysis, regression models are often specified using the hazard function and relationships are expressed using hazards ratio. However, in cases when the proportional hazards assumption is in question, it would be useful to be able to express the effect of covariates on a restricted mean survival time, in a manner similar to classical regression analysis which is focused on the mean of an outcome variable. Pseudo-values allow for this by replacing censored observations and event times with “leave-one-out” estimates at a given time (Andersen *et al*, 2003). Later, Andersen *et al* (2004) described the use of pseudo values as a route to assessing the effects of covariate(s) on restricted mean survival time. Royston and Parmar (2011) also provided a convincing argument for the use of a restricted mean when the proportional hazards assumption is in doubt. A restricted mean can be used where either the last observation is treated as an event or the investigator can assign an interval which is assumed to be the longest possible survival time for that study. Another version of the restricted mean is to assume the last event time as the last observed time regardless of later censored observations (Sheldon, 2006). In general, the choice of this point appears to be arbitrary and in all of the literature researched for this work, very little guidance is given or attention is paid to the choice of time point τ. Andersen *et al* (2004) performed simulation study for the choice of time point τ at 75th and 95th percentile of event time and reported that the biases are quite small for one of the choices. Therefore, to test the treatment effect with the presence of additional covariate(s), pseudo values regression approach based restricted mean survival time at 80th percentile of event time point was considered as an alternative and compared with new methods proposed by Callegaro *et al* (2014).

The restricted mean survival time θ_τ of a random variable T is the mean of $\min(T, \tau)$; it is the area under the survival curve $S(t)$ up to time τ and is given by:

$$\theta_\tau = E\big(\min(T,\tau)\big)$$
$$= \int_0^\tau S(t)\,dt$$

and can be estimated by:

$$\hat{\theta}_\tau = \int_0^\tau \hat{S}(t)\,dt$$

, where $\hat{S}(.)$ is the Kaplan Meier (1958) estimator and when T is the time to death, θ_τ might be interpreted as the τ year life expectancy (Royston *et al*, 2011). For a given restricted mean survival time point τ, let $\hat{S}_P(t)$ be pooled sample Kaplan–Meier estimator, based on all observations and $\hat{S}_P^{(j)}(t)$ be the Kaplan–Meier estimator based on the j^{th} observation removed. Then the j^{th} pseudo values restricted at time τ is defined by:

$\theta_{j\tau} = (n_c + n_E)\int_0^\tau \hat{S}_P(t)(t)dt - (n_c + n_E - 1)\int_0^\tau \hat{S}_P^{(j)}(t)dt$, j=1,2,...,n.

, where Kaplan–Meier (1958) estimator of survival in the k^{th} treatment group at event time t_j can be given as:

$$\hat{S}_k(t) = \prod_{t_j \le t}\left(1 - \frac{d_{kj}}{Y_{kj}}\right)$$

and its variance estimated by Greenwood's formula (Greenwood, 1926) has the form:

$$\widehat{Var}\left(\hat{S}_k(t)\right) = \left(\hat{S}_k(t)\right)^2 \sum_{t_j \le t}\left(\frac{d_{kj}}{Y_{kj} - d_{kj}}\right)$$

, where $t_1 < t_2 < \cdots < t_D$ are distinct event times, d_{kj} denote the number of events, and Y_{kj} denote the number of subjects at risk in the k^{th} treatment group at event time t_j and n_k is the number of subjects in the k^{th} treatment group for $k = 1$ for experimental (E) treatment and 0 for control (C) group. Once pseudo values are computed, then they can be used to model the effect of covariate(s) on the outcome (Klein *et al*, (2007), Logan *et al* (2008) and Andresen *et al* (2004)). The model based on these pseudo values restricted at time τ has the form:

$$g(\theta_{j\tau}) = \beta_0 + \beta_G G + \boldsymbol{X}'\boldsymbol{\beta}_{\boldsymbol{X}}, \text{ for } j = 1,2,...,n$$

, where G is treatment indicator (1 for experimental (E) treatment or 0 for control (C) group), X is vector of covariate(s) and $g(.)$ is the identity link function. Then, the null hypothesis of equality survival for patients in the treatment and control group is equivalent to testing $H_0: \beta_G = 0$ against one-sided alternative that experimental treatment increases survival time i.e. $H_A: \beta_G > 0$. Inference on β_G was performed using generalized estimating equations (Liang and Zegere, 1986) and the estimating equation to be solved has the form:

$$\sum_j \left(\frac{\partial}{\partial(\boldsymbol{\beta})} g^{-1}(\beta_0 + \beta_G G + X'\boldsymbol{\beta}_X)\right)' V_j^{-1}(\boldsymbol{\beta})\left(\theta_{j\tau} - \hat{\theta}_{j\tau}\right) = \sum_j U_j(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) = 0$$

, where $V_j(\boldsymbol{\beta})$ is a independence working covariance matrix, $\hat{\theta}_{j\tau}$ is the model based predicted values of $\theta_{j\tau}$. Let $\hat{\boldsymbol{\beta}}$ be the solutions to this equation then according Liang and Zeger (1986), under standard regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically multivariate normal with zero mean vector and covariance that can be estimated consistently by a "sandwich" estimator. Then the null hypothesis of no difference in survival times between treatment groups i.e. $H_0: \beta_G = 0$ against one-sided alternative that experimental treatment increases survival time i.e. $H_A: \beta_G > 0$ can be tested by:

$$Z_{PRMRST} = \frac{\hat{\beta}_G}{\sqrt{\widehat{var}(\hat{\beta}_G)}}$$

Under the null hypothesis, Z_{PRMRST} statistic assumed to follow a standard normal distribution for a large samples.

2.2.2. Testing for the Treatment Effect based on Extended Cox Model

Under the proportional hazards assumption, crossing of the survival curves is impossible. Thus, in a study where the patient groups do not differ between the treatments, crossing of the survival curves implies a violation of the proportional hazards assumption. If the proportional hazards assumption fails to hold for the treatment or for one or more of the covariates, the results of a Cox proportional hazards model will be misleading. Putter *et al* (2005) suggested a way of studying the effect of treatment changes over time by adding a time dependent treatment effects in a Cox proportional hazards model. The most straightforward way to model a time dependent treatment effect is by adding interaction terms of the treatment group with $f(t)$ as $\beta_G G(t) = \beta_G G f(t)$, where $f(t)$ is the function of time t with its popular choice can be t or log (t) or heaviside function that take value 1 for all time point greater than or equal to pre-specified time t_0 or zero otherwise. In this study, for the practicality and comparability of results, heaviside function which is conceptually related with stopped Cox and defined on the median of observed events (t_0) was adopted. In the literature it is stated that, if there is no information about crossing point for hazards the recommended choice is the time point where half of the expected number of event are observed (Zhou, 2006). Gillen and Emerson (2005) also suggested the use of equally spaced information time with the goal of balancing loss of statistical power against the potential for early stopping in the situation where there is no prior knowledge of a time varying treatment effect. These are considered as motivations for the choice of time point t_0 in this study. The general form of the extended Cox model with time dependent treatment effect can be written as:

$$h(t) = h_0(t)\exp\left(\beta_E G * (1 - \mathrm{f(t)}) + \beta_L G * \mathrm{f(t)} + \boldsymbol{X}'\boldsymbol{\beta_X}\right),\ \text{where}$$

$f(t) = \begin{cases} 1, & t \geq t_0 \\ 0, & t < t_0 \end{cases}$ is called heaviside (step) function, G is treatment groups (1 for treated and 0 for control), $\boldsymbol{X}$ are additional baseline covariate(s), β_E, β_L and $\boldsymbol{\beta_X}$ are parameters to be estimated representing early, late treatment effects and baseline covariate(s) effects, respectively. The parameters of the model were estimated by maximizing logarithm of partial likelihood via Newton-Raphson iterative procedure (Klein and Moeschberger, 1997). Lets denote $Pvalue_{early}$ and $Pvalue_{late}$ as one-sided p-value to test for the early and late treatment effect with hypothesis $H_{oe}\colon \beta_E = 0$ versus $H_{1e}\colon \beta_E < 0$ and $H_{ol}\colon \beta_L = 0$ versus $H_{1l}\colon \beta_L < 0$, respectively. Since early and late treatment effects are independent, the null hypothesis of $H_{oel}\colon H_{oe} \cap H_{ol}$ can be tested by combining two sub hypothesis by Fisher (1925) combining method as: $Z_{TD} = -2(\ln Pvalue_{early} + \ln Pvalue_{late})$ which is distributed chi-square with 4 degrees of freedom for two independent tests. The Z_{TD} tests if there is an early or a late treatment effect.

2.2.4. Testing for the Treatment Effect based on Combination of Treatment Effects from stopped/left truncated Cox and Cox proportional hazards Models

2.2.4.1. Cox Proportional hazards Model

In this section, the newly proposed methods by Callegaro *et al* (2014) to test for treatment effect based on Cox model, but stopped at different administratively censored time is described. It is well known that the proportional hazards model operates under the proportional hazards assumption, that the hazard for an individual in one treatment group at a given time is proportional to the hazard of a similar individual in the control group, and this proportion remains constant over time (Cox, 1972). Suppose that there is a total sample of n individuals with the survival time t and let G be treatment groups (1 for treated and 0 for control). Let $\boldsymbol{X}$ be set of additional baseline covariate(s), putting all of these elements together, the general form of the Cox (1972) proportional hazards model can be written as:

$$h(t) = h_0(t)\exp\left(\beta_G G + \boldsymbol{X}'\boldsymbol{\beta}_X\right)$$

, where $h_0(t)$ denote the hazard function for an individual on the control with covariate values all equal to zero, which is also known as the baseline hazard function. The parameters of the model were estimated by maximizing the logarithm of partial likelihood via Newton-Raphson iterative procedure (Klein and Moeschberger, 1997). From Cox proportional hazards model, the null hypothesis of no difference between treatments (i.e., $H_0\!: \beta_G = 0$) versus one-sided alternative that the treatment is better (i.e., $H_0\!: \beta_G < 0$) can be tested using:

$$Z_{Cox} = \frac{\hat{\beta}_G}{\sqrt{\widehat{var}(\hat{\beta}_G)}}$$

Under the null hypothesis, this statistic follows a standard normal distribution for a large samples. In the frame of Cox proportional hazards model, a Cox model stopped at t_0 is a Cox model fitted on the data with additional administrative censoring at time t_0 in order to study short term treatment effect. Van Houwelingen and Putter (2011) showed that the predictions based on the stopped Cox model are very accurate at the beginning of the follow-up and later in 2014 they concluded that stopped Cox works well for follow-up which is not too long. Furthermore, left truncated Cox model is also in the frame of Cox proportional hazards model fitted on the data left truncated at time t_0. To develop test statistics based on early effect or late effect and overall treatment effect, let $\hat{\beta}_E$ denote the treatment effect estimated by the stopped Cox model (early treatment effect), or $\hat{\beta}_L$ denote the treatment effect estimated by left truncated Cox (late treatment effect) or equivalently estimated from extended Cox model by using heaviside function and $\hat{\beta}_G$ represent overall treatment effect. According to Callegaro *et al* (2014), the effect of treatment can be tested by the sum of early (late) treatment effects and overall treatment effect and the test statistics has the form:

$$Z_{sumEO} = \frac{\hat{\beta}_E + \hat{\beta}_G}{\sqrt{\widehat{var}(\hat{\beta}_E) + 3\widehat{var}(\hat{\beta}_G)}}$$

or

$$Z_{sumLO} = \frac{\hat{\beta}_L + \hat{\beta}_G}{\sqrt{\widehat{var}(\hat{\beta}_L) + 3\widehat{var}(\hat{\beta}_G)}}$$

Under the null hypothesis, $Z_{sumEO}$and Z_{sumLO} statistics follow a standard normal distribution for large samples. Z_{sumEO} tests whether there is an overall or early treatment effects and Z_{sumEL} tests whether there is an overall or late treatment effects. They combines the two log hazards ratio by taking into account the dependence of the tests through covariance. Callegaro *et al* (2014) suggested to use the covariance between $\hat{\beta}_E(\hat{\beta}_L)$ and $\hat{\beta}_G$ as the variance of $\hat{\beta}_G$ and its theoretical derivation is related with theory of log-rank test (Mantel, 1966). In general, to compute Z_{sumEO} or Z_{sumLO} statistics, first the early, late and the overall treatment effects should be estimated in a way that the early and late treatment effect can be estimated by fitting the Cox model on data administratively censored at t_0 (the median of the observed event times) and left truncated Cox proportional hazards model after time t_0, respectively. Equivalently early and late treatment effects can be estimated from extended Cox model by using heaviside function. In this way, half of the events are used to estimate the early and late treatment effects. In general, the way to compute t_0 must be pre-specified in the protocol. The overall treatment effect can be estimated from Cox proportional hazards model. Another alternative is to combine two test statistics from early or late and the overall treatment effects using a group sequential like methodology. The global null hypothesis, that there is no treatment effect in the overall population (i.e., $H_{01}\colon \beta_G = 0$) nor in the subgroup (i.e., $H_{02}\colon \beta_E = 0$ or $\beta_L = 0$) is given by: $H_0\colon H_{01} \cap H_{02}$. The test statistics for group sequential like method have the form:

$$X_{GSEO} = Reject\ H_0\ if\ (Pvalue_{overall} < \alpha_1\ or\ Pvalue_{early} < \alpha_2) \text{ or}$$
$$X_{GSLO} = Reject\ H_0\ if(Pvalue_{overall} < \alpha_1\ or\ Pvalue_{late} < \alpha_2)$$

, where $Pvalue_{overall}$, $Pvalue_{early}$ and $Pvalue_{late}$ are p-values from overall, early and late treatment effects, respectively. The significance levels are denoted by α_1 and α_2. To control the family wise error rate below a value α for a pre-specified significance level α_1, α_2

is defined in such a way that prob$\left(Z > Z_{\alpha_1}\ or\ Z(t_0) > Z_{\alpha_2}|H_0\right) = \alpha$. Spiessens and Debois (2010) showed that α_2 can be determined by solving the equation:

$$\int_{-\infty}^{Z_{\alpha_1}} \Phi\left(\frac{Z_{\alpha_2} - \sqrt{\tau}\,Z}{\sqrt{1-\tau}}\right)\Phi(Z)d_Z = 1 - \alpha$$

, where τ is the information fraction in the subgroup and is given by: $\hat{\tau} = \frac{v\widehat{ar}(\hat{\beta}_T)}{v\widehat{ar}(\hat{\beta}_T(t_0))}$. The level of significance α_1 was used in group sequential method under overall treatment effect from Cox proportional hazards model and α_2 was used for early or late treatment effects. For administratively censored time point t_0 for which about half of the observed events are censored $i.e., \hat{\tau} = 0.5$ and for fixed $\alpha_1 = 0.03$ significance level, α_2 was calculated to be 0.017 which was computed by using standard package for group sequential design in R. In general, test statistics from group sequential like method i.e., X_{GSEO} or X_{GSLO} combines the two p-values from early and overall treatment effects or late and overall treatment effects, respectively. This method takes dependence of the tests into account by group-sequential like approach i.e., by splitting significance level.

Finally, another proposed test statistic was to choose the time point t_0 which maximizes the treatment effect of the extended Cox model and has the form:

$$Z_{max} = Max_{t_0}\{Z_{TD}(t_0)\} = Max_{t_0}\left\{-2(\ln Pvalue_{early}(t_0) + \ln Pvalue_{late}(t_0))\right\}$$

, where Z_{max} is test statistic in which the maximum of the treatment effect is observed. The treatment effect is estimated by fitting the extended Cox model with heaviside function at the event time t_k, for k= 1,..., K. The Z_{max} statistic is the maximum of the K estimated treatment effects and its distribution under the null hypothesis is not known. In this case a permutation test, where the treatment label is permuted was used to derive the p-value. In order to perform the permutation test, compute test statistic for the actual data (Z_{max}) from event time t_k, for k= 1,..., K and calculate the values of the same statistic for each of the possible assignments of the treatment labels of the total n observations by permuting treatment label. Finally, the proportion of these values that are equal to or greater than the value of the statistic for the actual data is the desired p-value. In this study, due to computational intensive nature of the test, 300 random possible arrangements of the treatment label was adopted.

3. RESULTS

3.1. Simulation Results

In order to evaluate the performance of newly proposed methods for testing the effect of treatment in randomized clinical trial when proportional hazards assumption is in doubt, survival data was simulated from a population exhibiting different possible treatment effects as displayed in Figure 1.

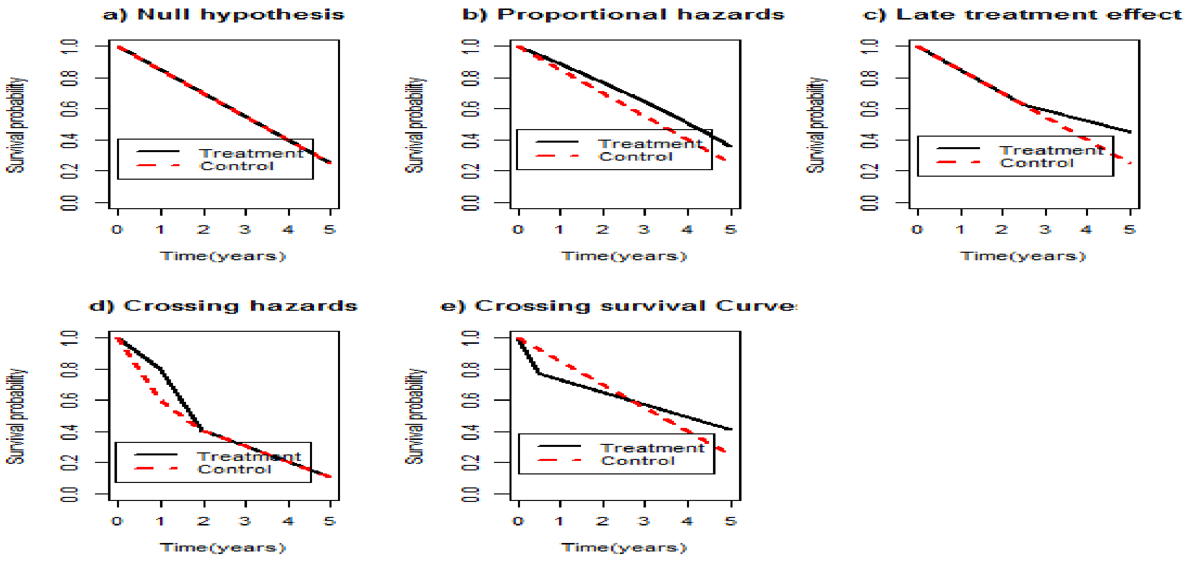

Figure 1: T*rue survival curves used to simulate the data under different scenarios*

Figure 1 displays true survival curves that were used to simulate sample data under five possible treatment effects such as: a) no treatment effect (under the null), b) constant beneficial effect (proportional hazards alternative), c) no initial effect but a gradually increasing beneficial effect, d) an initial beneficial effect that diminishes long-term and e) an initial harmful and late beneficial effect of treatment. The sample data was replicated 1000 times under different scenarios containing a total of 400 subjects with one to one randomization and 30% administrative censoring. The average estimated treatment effects from the simulated data with and without covariate in the models is presented in the Table A_1 and A_2 in the Appendix.

Table 1: Estimated type I error rate and Power of the tests based on the simulations under different scenarios for different methods without including covariate in the models

	Methods							
Under	Z_{sumEO}	Z_{sumLO}	X_{GSEO}	X_{GSLO}	Z_{max}	Z_{Cox}	Z_{PRMRST}	Z_{TD}
Null Hypothesis	0.059	0.055	0.059	0.055	0.050	0.054	0.057	0.058
H_A of PH Assumption	0.724	0.514	0.701	0.658	0.753	0.743	0.679	0.714
H_A of Late Treatment	0.670	0.943	0.831	0.922	0.979	0.892	0.405	0.909
H_A of Early Treatment	0.872	0.005	0.954	0.207	0.948	0.266	0.933	0.896
H_A of Crossing Survivals	0.236	0.779	0.444	0.741	0.939	0.540	0.030	0.675

Z_{Cox} represents test statistic from Cox proportional hazards model, $Z_{sumEO}(Z_{sumLO})$ is test based on sum of early and overall treatment effects(sum of late and overall treatment effects), X_{GSEO} (X_{GSLO}) is from group sequential method based on test statistics from early and overall treatment effects (late treatment and overall treatment effects), Z_{max} is maximum of test statistics from all distinct event time points by permutation, Z_{PRMRST} is from pseudo values regression approach based on restricted mean survival time and Z_{TD} *is from extended Cox model.*

Table 1 displays simulation results of the estimated type I error rate and power of newly proposed tests, test from Cox proportional hazards, pseudo values regression approach based on restricted mean survival time and extended Cox without including the covariate in the models. From Table 1, it can be observed that under the null hypothesis all methods controlled a type I error rate stabilizing around the targeted 0.05 level of significance and this was expected in order for the test method to be efficient. The power of the various procedures is expected to depend heavily on the scenarios, for instance, the test from Cox proportional hazards model is expected to perform well in case of proportional hazards alternative. However, it can be seen that the newly proposed tests as well as test from extended Cox model performed about the same as compared to the test from Cox proportional hazards model. Under proportional hazards alternative, the test from pseudo values regression approach based on restricted mean survival time had less power compared to test from Cox proportional hazards model. It was also seen that in the case of late treatment effect, the test from Cox proportional hazards model, test statistics based on the sum of late and overall treatment effects, group sequential like method based on late and overall treatment effects, permutation test based on maximum treatment effect and test from extended Cox model performed reasonably well under this scenario. As was expected, in the situation where two survival curves are separate at the beginning and then close as time goes on (i.e., early treatment effect) and crossing survival curves, the tests for treatment effect from Cox proportional hazards model had less power and this might be due to the contrasting early and late effects of the treatments. From newly proposed methods, test statistics based on the sum of early and overall treatment effects, group sequential like method based on early and overall treatment effects, permutation test based on maximum treatment effect had better performance under early treatment effect in which hazards are expected to cross. They performed about the same compared to pseudo values regression approach based on restricted mean survival time and extended Cox model under this scenario. On the other hand, test statistics based on the sum of late and overall treatment effects, group sequential like method based on late and overall treatment effects, permutation test based on maximum treatment effect had better performance under the crossing survival curves. They performed about the same compared to test from extended Cox model and better compared to pseudo values regression approach based on restricted mean survival time. Overall, from new methods, permutation test statistic showed better performance under all alternative scenarios although it is computational intensive.

Table 2: Estimated type I error rate and power of the tests based on the simulations under different scenarios for different methods with including covariate in the models

	Methods							
Under	Z_{sumEO}	Z_{sumLO}	X_{GSEO}	X_{GSLO}	Z_{max}	Z_{Cox}	Z_{PRMRST}	Z_{TD}
Null Hypothesis	0.058	0.050	0.055	0.051	0.050	0.059	0.058	0.057
H_A of PH Assumption	0.728	0.721	0.757	0.747	0.739	0.800	0.646	0.771
H_A of Late Treatment	0.306	0.948	0.757	0.942	0.982	0.830	0.103	0.915
H_A of Early Treatment	0.977	0.013	0.997	0.402	0.983	0.495	0.994	0.985
H_A of Crossing Survivals	0.006	0.885	0.224	0.970	0.964	0.311	0.001	0.937

Z_{Cox} represents test statistic from Cox proportional hazards model, $Z_{sumEO}(Z_{sumLO})$ is test based on sum of early and overall treatment effects(sum of late and overall treatment effects), X_{GSEO} (X_{GSLO}) is from group sequential method based on test statistics from early and overall treatment effects (late treatment and overall treatment effects), Z_{max} is maximum of test statistics from all distinct event time points by permutation, Z_{PRMRST} is from pseudo values regression approach based on restricted mean survival time and Z_{TD} *is from extended Cox model.*

Table 2 displays simulation results of the estimated type I error rate and power of newly proposed tests, test from Cox proportional hazards model, pseudo values regression approach based on restricted mean survival time and extended Cox with the presence of covariate in the models. In general, when covariate is introduced into the models the pattern of results in terms of maintaining type I error and the power of the tests was similar to the results obtained without covariate in the models (Table 1). However, there was a gain in power for most of methods when covariate is included in the models. Specifically, in contrast to the Cox proportional hazards model, test statistics based on sum of early and overall treatment effect, group sequential like method based on early and overall and permutation test in which the effect of treatment maximized were powerful in the case of early treatment effect where hazards are expected to cross. They also perform similarly compared to pseudo values regression approach based on restricted mean survival time and extended Cox model under this scenario.

On the other hand, test statistics based on the sum of late and overall treatment effects, group sequential like method based on late and overall treatment effect and permutation test were powerful in the case of crossing survival curves whereby there is an initial harmful and late beneficial effects of the experimental treatment. Furthermore, in case of proportional hazards alternative, using the newly proposed methods did not yield dramatic decrease in statistical power compared to the Cox proportional hazards model.

3.2. Implementation of the Methods on the Real Datasets

To evaluate the performance of newly proposed methods, two dataset on crossing survival curves and late treatment effects were analyzed and results are displayed in the subsequent sections. The detail description about dataset is given in section 2.1. Kaplan Meier survival curves were used as an exploratory tool in order to describe the data.

3.2.1. Gastric cancer dataset

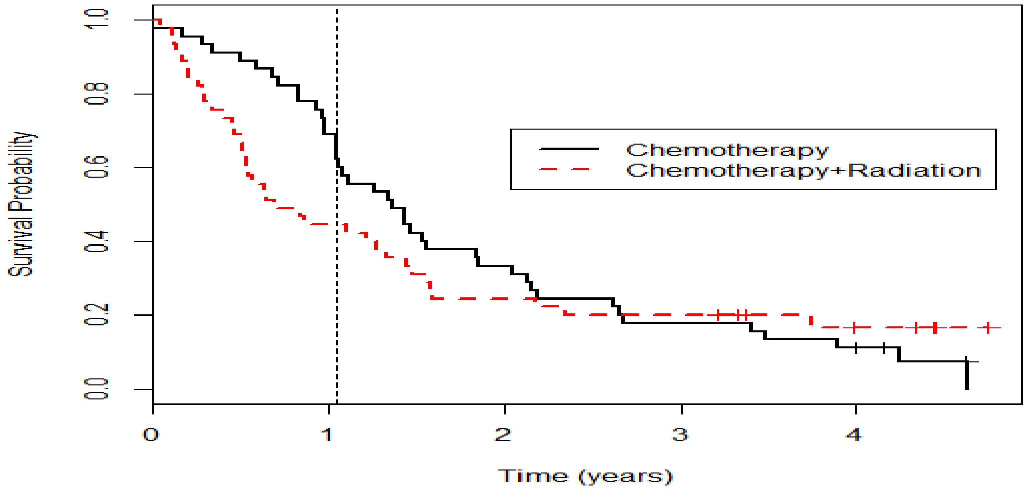

Figure 2: Kaplan-Meier estimated survival curves for the gastric cancer data-set by treatment groups

Figure 2 displays Kaplan Meier plots of overall survival curves by treatment group. Clearly from Figure2, it can be seen that the treatment effect (chemotherapy plus radiation) was initially unfavorable and later became advantageous over control (chemotherapy without radiation). The two curves of the treatment group crossed after about 2.5 years. From log-log survival plot in Figure 3, it can be seen that two survival curves are not parallel. The crossing survival curves and lack of parallelism on log-log plot are a clear sign of non-proportionality in which Cox proportional hazards model might not work well. The dashed vertical lines on the plot represent the medians of the observed event time point ($t_0 = 1.04$).

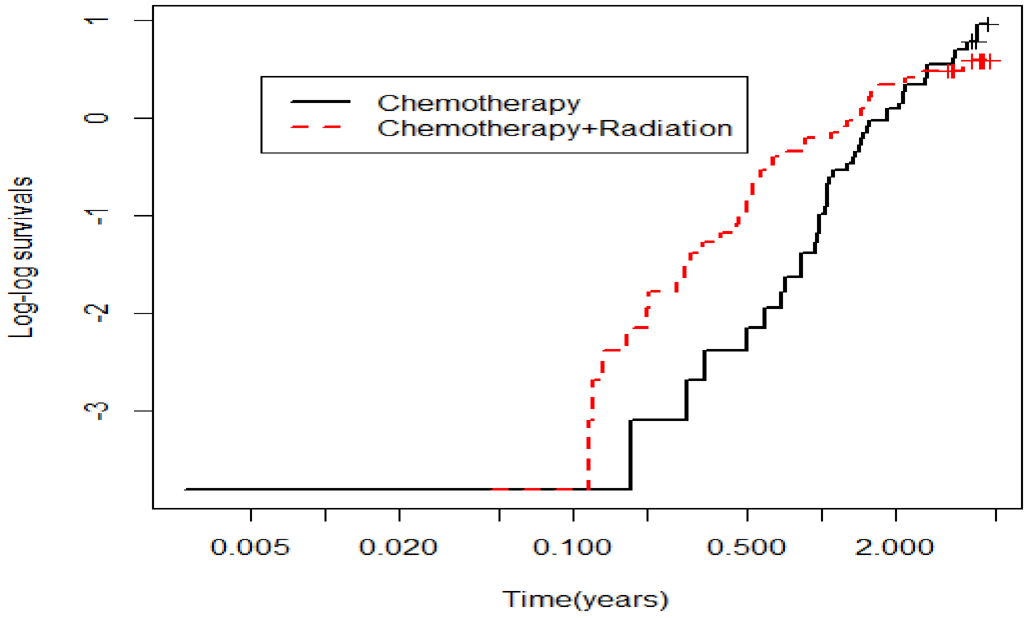

Figure 3: Log-log survival plot for gastric cancer dataset by treatment groups

Table 3: P-values from one-sided test statistics to test treatment effect under crossing survival curves

	Methods							
	Z_{sumEO}	Z_{sumLO}	X_{GSEO}	X_{GSLO}	Z_{max}	Z_{Cox}	Z_{PRMRST}	Z_{TD}
P-values	0.976	0.204	0.993/0.733	0.049/0.733	0.197	0.733	0.995	0.197

Z_{Cox} represents test statistic from Cox proportional hazards model, $Z_{sumEO}(Z_{sumLO})$ is test based on sum of early and overall treatment effects(sum of late and overall treatment effects), Z_{max} is maximum of test statistics from all distinct event time points by permutation, Z_{PRMRST} is from pseudo values regression approach based on restricted mean survival time , *Z_{TD} is from extended Cox model. and P-values reported for X_{GSEO} and X_{GSLO} are from early/overall and late/overall treatment effects, from group sequential like method, respectively.*

Table 3 shows one-sided p-values of the test statistics from new proposed methods, Cox proportional hazards model, pseudo values regression approach based on restricted mean survival time and extended Cox model to test for the effect of treatment. From the results, newly proposed test statistics based on sum of late and overall treatment effects and permutation test performed about the same compared to extended Cox, but better than test from Cox proportional hazards model and pseudo values regression approach based on restricted mean survival time. Moreover, there was late beneficial treatment effect as p-value from the test statistic from group sequential like method at late treatment was small as compared to test statistic based on early treatment effect although it was statistically insignificant at 2% level of significance from group sequential like method to test for early (late) treatment effects (i.e., 0.049 >0.02). This was also reflected through one-sided p-values from test statistic based on the sum of late and overall treatment effects. These results are consistent with the findings of the simulation studies under crossing survival curves.

3.2.2. Bladder cancer dataset

In order to describe survival distribution of treatment groups for bladder cancer dataset, Kaplan Meier survival curves were used as presented in Figure 4.

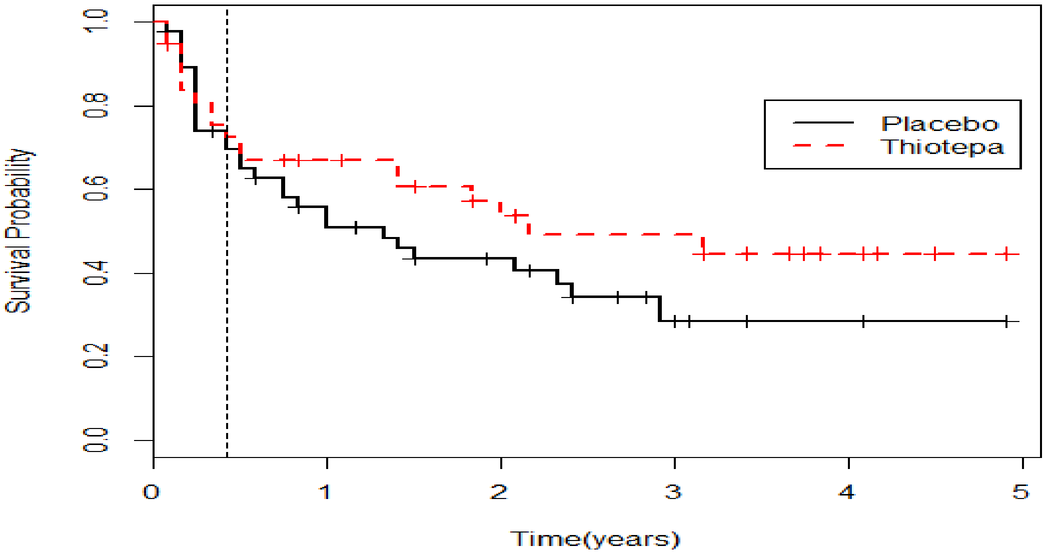

Figure 4: Kaplan-Meier survival curves for the bladder cancer dataset by treatment groups

Figure 4 displays the plots of Kaplan Meier estimated survival probabilities by treatment groups. From the figure, it can be seen that there was delay in the effect of the treatment as Kaplan Meier survival curves are start to diverge after half of the observed event time point. The dashed vertical lines on the plot represent the medians of the observed event time point ($t_0 = 0.41$) and clearly the data exemplify the late treatment effect.

Table 4: Parameter estimates (standard errors) obtained from Cox proportional hazards model and stopped Cox at 0.41 years under late treatment effect

	Cox PH Model		Stopped Cox Model	
Effects	Coeff. (se. coeff)	P-values	Coeff. (se. coeff)	P-values
Treatment	-0.5260 (0.3158)	0.0479*	-0.2351 (0.465)	0.3067
Initial number of tumors	0.2382 (0.0759)	0.0017*	0.2403 (0.104)	0.0210*
Size of tumors	0.0696 (0.1016)	0.4900	0.0441 (0.155)	0.780

* Statistically significant at 5% level of significance, Coeff. represent estimated parameters, se. coeff is standard errors of estimated parameters.

Table 4 displays the parameter estimates (standard errors) and their corresponding p-values from Cox proportional hazards and stopped Cox models. From the Table 4, the initial number of tumors was significantly associated with death among bladder cancer patients. The effect of treatment was statistically not significant in stopped Cox (one-sided pvalue=0.3067), but borderline significant in Cox proportional hazards model (one-sided p-value=0.0479). The treatment effect stopped at 0.41 years was lower than the overall treatment effect from Cox proportional hazards model with higher standard error. As was expected from the simulation results under late treatment effect, the p-value of the classical Cox model is smaller than the p-value of the stopped Cox model.

Table 5: Parameter estimates (standard errors) obtained from regression approach based on restricted mean survival time at 80% of observed event time under late treatment effect

Effects	Coeff.	SE.coeff.	P-values
Intercept	1.1516	0.1572	<0.0001*
Treatment	0.1348	0.1169	0.1244
Initial number of tumors	-0.0890	0.0392	0.0130*
Size of tumor	-0.0236	0.0473	0.5950

** Statistically significant at 5% level of significance,* Coeff. represent estimated parameters, se. coeff. is standard errors of estimated parameters.

Table 5 shows estimated parameters and their standard errors from pseudo values regression approach based on mean restricted survival time on the 80% event observed time point. From the results, it was seen that survival time of patients significantly related with the initial number of tumors. So, for a unit increase in initial number of tumors, the mean restricted survival time of the patients decrease by 0.089 years. Comparing results from stopped Cox and pseudo values regression approach based on the restricted mean survival time with Cox proportional hazards model, they produced higher one-sided p-values for the treatment effect. This was not surprising as it was evident from simulation results that the Cox proportional hazards model works reasonable well under late treatment effect.

Table 6: Parameter estimates (standard errors) obtained from extended Cox model under late treatment effect

Effects	Coeff.	SE.coeff.	P-values
Early Treatment	-0.2696	0.4269	0.2638
Late Treatment	-0.7966	0.4513	0.0388*
Initial number of tumors	0.2351	0.0760	0.0020*
Size of tumor	0.0735	0.1014	0.4682

** Statistically significant at 5% level of significance,* Coeff. represent estimated parameters, se. coeff. is standard errors of estimated parameters.

Table 6 displays the parameter estimates (standard errors) obtained from extended Cox model. By combining one sided p-values of early and late treatment effect, the p-value from extended Cox model was found to be 0.057 which is borderline significant. However, the risk of dying was significantly lower for patients in the treatment group compared to control group after the median of observed event time point (one-sided p-value=0.038). As before, the initial number of tumors had statistically significant effect on the risk of dying. Moreover, to illustrate the performance of newly proposed methods compared to Cox proportional hazards model, pseudo values regression approach based on restricted mean survival time and extended Cox model, one sided p-values of the test statistics are given in Table 7.

Table 7: P-values from one-sided test statistics to test for treatment effect with and without including the covariates in the models under late treatment effect

Without including the covariates in the models								
	Z_{sumEO}	Z_{sumLO}	X_{GSEO}	X_{GSLO}	Z_{max}	Z_{Cox}	Z_{PRMRST}	Z_{TD}
P-values	0.246	0.064	0.415/0.110	0.065/0.110	0.116	0.110	0.193	0.124
With including the covariates in the models								
P-values	0.126	0.0310*	0.264/0.048	0.039/0.048	0.053	0.048	0.124	0.057*

Z_{Cox} represents test statistic from Cox proportional hazards model, $Z_{sumEO}(Z_{sumLO})$ is test based on sum of early and overall treatment effects(sum of late and overall treatment effects), Z_{max} is maximum of test statistics from all distinct event time points by permutation, Z_{PRMRST} is from pseudo values regression approach based on restricted mean survival time , Z_{TD} *is from extended Cox model. and P-values reported for* X_{GSEO} *and* X_{GSLO} *are from early/overall and late/overall treatment effects from group sequential like method, respectively.*

Table 7 displays one-sided p-values from newly proposed methods, Cox proportional hazards model, pseudo values regression approach based on restricted mean survival time and extended Cox with and without including covariates in the models. From the Table 7, in the presence of covariates in the models, the effect of treatment was borderline significant in the Cox proportional hazards model, extended Cox and permutation test. As was expected, from new proposed methods, test statistic based on sum of late and overall treatment effects and permutation test performed about the same compared to the test from Cox proportional hazards and extended Cox models, but perform better compared to pseudo values regression approach based on restricted mean survival time. However, test statistics based on sum of early and overall treatment effects had less power compared to Cox proportional hazards and extended Cox models. These results are consistent with the findings of the simulation study under late treatment effect.

4. DISCUSSION AND CONCLUSIONS

The Cox proportional hazards model is the standard approach to evaluate the treatment effect on clinical trial data. When non-proportional hazards is present Cox model may not be powerful, especially in the case of crossing hazards. In such a case, the test for the difference in hazard rates between the treatments will most likely not be significant, because of the contrasting early and late effects of the treatments. Different approaches have been proposed as alternative to the Cox proportional hazards model in the case of non-proportional hazards. Therefore, the main purpose of this thesis was to evaluate the performance of one sided newly proposed methods by Callegaro *et al* (2014) for testing the treatment effect in randomized clinical trials when proportional hazards assumption is in doubt. They were compared with Cox proportional hazards model, pseudo values regression approach based on restricted mean survival time and extended Cox model. This was done based on simulations and two popular real datasets exhibiting crossing survivals curves and late treatment effect. Performance of new proposed methods were evaluated in terms of maintaining nominal level of significance and empirical power. From simulation results, it was seen that all methods controlled the type I error rate accurately in a sense that empirical type I errors were close to the targeted 0.05 level of significance with and without including covariates in the models. Hence, the normal distribution seems an adequate approximation for the sample sizes investigated. As was expected, the performance of the Cox proportional hazards model for testing treatment effect generally lacks power in situations where there is early treatment effect and two survival curve cross. Simulation results showed that the newly proposed methods of testing treatment effect; test statistics based on sum of early and overall treatment effects, group sequential tests based on early and overall, and permutation test based on maximum treatment effect performed reasonably well compared to Cox proportional hazards model under early treatment effect where hazards are expected to cross. They also performed about the same compared to pseudo values regression approach based on restricted mean survival time at 80% of the observed event time and extended Cox model in the case of early treatment effect. It was seen that permutation test had better results under four alternative scenarios compared to the power of other newly proposed test statistics. These results are similar to the finding by Callegaro *et al* (2014).

In general, using the newly proposed methods under proportional hazards alternative did not yield decreases in statistical power compared to the Cox proportional hazards model, pseudo values regression approach based on restricted mean survival time and extended Cox model.

It should be noted that the performance of test statistics based on sum of early (late) and overall treatment effects, group sequential like method based on early (late) and overall treatment effects, pseudo values regression approach based on restricted mean survival time and extended Cox depends on choice of time points. Hence, the way to compute time point t_0 must be pre-specified in the protocol. The advantage of the test statistic based on maximum treatment evaluated at all event times with respect to the other test statistics is that its results do not depend on a pre-specified time point t_0. However, its drawback is that the distribution of the test statistic is unknown. Hence, a permutation test was used to compute the p-value which is computational intensive. As indicated in the simulation studies, new methods proposed by Callegaro *et al* (2014) reject the null hypothesis if a beneficial treatment effect is observed at a certain time point, irrespective of possible harmful treatment effects observed at other time points. In conclusion, new proposed methods are straightforward to implement in most statistical packages and allow to adjust for covariates as they performed reasonable well with the presence of covariate(s) in the models. They are useful for testing the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. The proposed methods can be particularly useful in cancer clinical trials with long-term follow-up as they are powerful in case of crossing survival curves whereby there is an initial harmful and late beneficial effects of the experimental treatment.

There was a few limitation to this simulation studies. This include: the study considered simulation setting for a sample of 200 subjects per treatment group with 30% administrative censoring, in the future work, one can investigate the different censoring rates and sample size effects to see how that would directly affect the results of the power and type I error rate of the newly proposed test statistics.

5. REFERENCES

Andersen, P.K, Hansen, M.G and Klein, J.P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis*; 10:335-350.

Andersen, P.K, Klein, J.P and Rosthoj, S. (2003). Generalized linear models for correlated pseudo-observations with applications to multi-state models. Biometrika;90:15–27.

Byar, D.P. (1984). The Veterans Administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine and topical thiotepa. In Bladder Tumors and Other Topics in Urological Oncology, (Edited by m. Pavone-Macaluso, P. H. Smith and F. Edsmyr). Plenum, New York, 363-370.

Burton, A, Altman, D.G, Royston, P, and Holder, R.L. (2006). The design of simulation studies in medical statistics: Wiley InterScience, *Statist. Med,* 25:4279–4292

Bain, L and Engelhardt, M. (1991). Statistical Analysis of Reliability and Life testing Models: Theory and Methods. Marcel Dekker, Inc., New York, 2nd edition.

Callegaro, A, Debois, M and Spiessens, B. (2014). Testing the treatment effect in randomized clinical trials with possible non-proportional hazards: Working paper: GSK Vaccines, Belgium.

Chen, P and Tsiatis, A.A. (2001) . Causal inference on the difference of the restricted mean lifetime between two groups, Biometrics vol. 57 pp. 1030–1038.

Cox, D.R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society, Series B, 34 (2), 187-220.

Fisher, R.A. (1925). Statistical Methods for Research Workers. Oliver and Boyd (Edinburgh). ISBN 0-05-002170-2.

Gastrointestinal Tumor Study Group. (1982). A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. Cancer.

Gillen, D.L and Emerson, S.S. (2005). A note on P-Values under Group Sequential Testing and Non proportional Hazards: Biometrics 61, 546-551, *DOI: 10.1111/j.1541-0420.2005.040342.x*

Greenwood, M. (1926) The natural duration of cancer , in *Reports on Public Health and Medical Subjects* , vol. 33 , Her Majesty ' s Stationary Office , London , pp. 1 – 26 .

Kaplan, E.L and Meier, P. (1958). Nonparametric estimation from incomplete observations . *Journal of the American Statistical Association* , 53 , 457– 481 .

Klein, J.P, Logan, B.R and Harhoff, M and Andersens, P.K. (2007). Analyzing survival curves at a fixed point in time. Statistics in Medicine, 26, 4505-4519.

Klein, J. P. and Moeschberger, M. L. (1997). Survival Analysis: Techniques for Censored and Truncated Data, New York: Springer-Verlag.

Liang, K.Y and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Logan, B.R, Klein, J.P and Zhang, M.J. (2008). Comparing Treatments in the Presence of Crossing Survival Curves: An Application to Bone Marrow Transplantation: *Biometrics: 733–740. doi:10.1111/j.1541-0420.2007.00975.x.*

MacKenzie, G and Ha, Li Do. (2007). Modelling Survival Data with Crossing Hazards. IWSM, P3-4.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50** (3): 163–70. PMID 5910392: can be accessed: http://en.wikipedia.org/wiki/Log-rank_test

Oquigley, J and Pessione, F. (1991). The problem of a covariate time qualitative interaction in a survival study. Biometrics; 47:101–115.

Putter, H, Sasako, M, Hartgrink, H.H, van de, Velde C.J and van Houwelingen, J.C. (2005). Long-term survival with non-proportional hazards: results from the Dutch Gastric Cancer Trial. Stat Med, 24, 2807-2821.

Royston, P. and Parmar, M.K.B. (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. Statist. Med., 30, 2409-2421.

Schemper, M., Wakounig, S., and Heinze, G. (2009). The estimation of average hazard ratios by weighted Cox regression. Statist. Med., 28, 2473-2489.

Sheldon, E.H. (2006). Choosing the Cut Point for a Restricted Mean in Survival Analysis, a Data Driven Method, PhD dissertation at Virginia Commonwealth University.

Spiessens, B. and Debois, M. (2010). Adjusted significance levels for subgroup analysis in clinical trials. Cont Clin Trials, 31, 647-656.

Stablein, D. M., Carter, W. H., and Novak, J.W. (1981). Analysis of survival data with non proportional hazards functions. Controlled Clinical Trials, 2, 149-159.

Yang, S and Zhao, Y. (2007). Testing treatment effect by combining weighted log-rank tests and using empirical likelihood: Science Direct: Statistics & Probability Letters 77: 1385–1393.

Van Houwelingen, H.C, and Putter, H. (2014). Comparison of stopped Cox regression with direct methods such as pseudo-values and binomial regression. Lifetime Data Anal, DOI 10.1007/s10985-014-9299-3.

Wei, L.J, Lin, D.Y and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84.

Zhang, M.J and Klein, J.P. (1998). Confidence Bands for the Difference of Two Survival Curves Under Proportional hazards Model: *Technical Report 29, Medical College of Wisconsin*

Zhou, M. (2006). Log-rank Test: When does it Fail and how to fix it: University of Kentucky, Department of Statistics Tech Report: Submitted/under revision: http://www.ms.uky.edu/~mai/research/LogRank2006.pdf accessed on July 25, 2014.

6. Appendix

Table A$_1$: Estimated treatment effects based on the simulations under different scenarios with 30% administrative censoring for a sample of size 200 per group without including the covariate in the models

	Treatment effects(log-hazards ratio)					
Under	β_G	β_E	β_L	$\beta_G+\beta_E$	$\beta_G+\beta_L$	$\beta_G(PRMST)$
Null Hypothesis	-0.001	-0.004	0.001	-0.005	-0.000	0.003
H_A of PH assumption	-0.249	-0.305	-0.193	-0.553	-0.442	0.316
H_A of Late Treatment	-0.269	-0.201	-0.337	-0.469	-0.606	0.244
H_A of Early Treatment	-0.122	-0.629	0.388	-0.751	0.267	0.195
H_A of crossing survivals	-0.178	-0.016	-0.345	-0.194	-0.524	-0.047

β_G represents overall from Cox PH model, β_E and β_L are is early and late treatment effects from extended Cox, $\beta_G+\beta_E$ is sum of early and overall treatment effects from Cox and extended Cox, $\beta_G+\beta_L$ is sum of late and overall treatment effects from Cox and extended Cox and $\beta_G(PRMST)$ is overall treatment effects from pseudo values regression approach based on restricted mean survival time.

Table A$_2$: Estimated treatment effects based on the simulations under different scenarios with 30% administrative censoring for a sample of size 200 per group with including covariate in the models

	Treatment effects (log-hazards ratio)					
Under	β_G	β_E	β_L	$\beta_G+\beta_E$	$\beta_G+\beta_L$	$\beta_G(PRMST)$
Null Hypothesis	-0.002	-0.006	0.001	-0.008	-0.001	0.003
H_A of PH assumption	-0.305	-0.306	-0.304	-0.611	-0.609	0.231
H_A of Late Treatment	-0.307	-0.001	-0.625	-0.309	-0.932	0.045
H_A of Early Treatment	-0.197	-0.798	0.409	-0.995	0.212	0.954
H_A of crossing survivals	-0.126	0.415	-0.690	0.288	-0.816	-0.243

β_G represents overall from Cox PH model, β_E and β_L are is early and late treatment effects from extended Cox, $\beta_G+\beta_E$ is sum of early and overall treatment effects from Cox and extended Cox, $\beta_G+\beta_L$ is sum of late and overall treatment effects from Cox and extended Cox and $\beta_G(PRMST)$ is overall treatment effects from pseudo values regression approach based on restricted mean survival time.

Table A$_3$: R-code to simulated sample data under five different scenarios conditioning on the covariates

```
#########################################################################################
# Function to simulate Sample data from the true survival curves under different Scenarios
#########################################################################################
simulate.data<-function(n,prop.obs=0.7,config=0, beta.z=0.5){
time=seq(0,5,length=1000)
z=rbinom(n, size=1, prob=0.5)
#########################################################################################
 # Survival functions under null hypothesis
#########################################################################################
if (config==0) surv1=surv0=1-.15*time
####################################################################### ##################
 # Survival functions under Proportional assumption
####################################################################### ##### ###### ####
if (config==1) {
surv1=(1-.15*time)^exp(-0.3);
surv0=(1-.15*time);
}
#########################################################################################
 #Survivalfunctions under the late treatment effect alternative
#########################################################################################
```

```
if (config==2) {
surv1=surv0=numeric()
sel=time<=2.5
surv1[sel]=surv0[sel]=(1-.15*time[sel])
sel=time>2.5
surv1[sel]=(.8-.07*time[sel])
surv0[sel]=(1-.15*time[sel])
}
########################################################################################
 #Survival functions under Crossing hazards/Early treatment Effect
########################################################################################
if (config==3) {
surv1=surv0=numeric()
 surv0[time<1]=1-.4*time[time<1]
 surv0[time>1 & time<2]=.8-.2*time[time>1 & time<2]
 surv0[time>=2]=0.605-0.1*time[time>=2]
 surv1[time<1]=1-.2*time[time<1]
 surv1[time>1 & time<2]=1.2-.4*time[time>1 & time<2]
 surv1[time>=2]=0.605-0.1*time[time>=2]

}
#########################################################################################
 #Survival functions under Crossing Survival Curves
#########################################################################################
if (config==4) {
surv1=surv0=numeric()
tt0=0.5
sel=time<=tt0
surv0[sel]=(1-.15*time[sel])
surv1[sel]=(1-.45*time[sel])
sel=time>tt0
surv1[sel]=(.81-.08*time[sel])
surv0[sel]=(1-.15*time[sel])
}
#########################################################################################
# To generate random variable from uniform to simulate random survival times and covariates
#########################################################################################
z=rbinom(n, size=1, prob=0.5)
surv1.z=surv1^exp(beta.z)
surv0.z=surv0^exp(beta.z)
u1=runif(n);u0=runif(n)
t1=t0=numeric()
######################################################################################
# Kind of Inversion Method to generate random survival times conditioning on covariate
######################################################################################
for(i in 1:n){
if(z[i]==0){
        t1[i]=max(time[abs(surv1-u1[i])==min(abs(surv1-u1[i]))])
        t0[i]=max(time[abs(surv0-u0[i])==min(abs(surv0-u0[i]))])
        }
if(z[i]==1){
        t1[i]=max(time[abs(surv1.z-u1[i])==min(abs(surv1.z-u1[i]))])
        t0[i]=max(time[abs(surv0.z-u0[i])==min(abs(surv0.z-u0[i]))])
        }
}
t=c(t1,t0)
########################################################################################
# For administered censoring (censoring due to end of the study)
#######################################################################################
thr=quantile(t,prob=prop.obs)
```

```
status=numeric()
status[t>thr]=0
status[t<=thr]=1
tt=t
tt[t>thr]=thr
oo=data.frame(tt, status, c(rep(1,n),rep(0,n)),z)
colnames(oo)=c('time','event','G','z')
oo
}
#############################################################################################
#Functions for Group Sequantial design to compute alphas and for other methods
#############################################################################################
find.alpha2.twosided<-function(alpha1,alpha,tau, interval=c(-3, 3)){
f<-function(z1,z.alpha2,alpha,alpha1,tau){
z=(z.alpha2-sqrt(tau)*z1)/sqrt(1-tau)
pnorm(z)*dnorm(z1)
}
f2<-function(z.alpha2,alpha1,alpha,tau){
integrate(f,lower=-Inf, upper=qnorm(1-alpha1),z.alpha2=z.alpha2,alpha=alpha,alpha1=alpha1,tau=tau)
}
f3<-function(z.alpha2,alpha1,alpha,tau){
f2(z.alpha2=z.alpha2,alpha=alpha,alpha1=alpha1,tau=tau)$value-(1-alpha)
}
alpha1=alpha1/2
alpha=alpha/2
z.alpha2=uniroot(f3,interval,alpha1=alpha1,alpha=alpha,tau=tau)$root
2*(1-pnorm(z.alpha2))
}
############################################################################################
#Function to apply Cox model on censored data-set (one-sided)
############################################################################################
stopped.cox<-function(data,t0){
data0=data
sel=data0$time>=t0
data0$event[sel]=0
data0$time[sel]=t0
cc=coxph(Surv(time, event)~G+z, data0)
beta=coefficients(summary(cc))[1,1]
sd.beta=coefficients(summary(cc))[1,3]
pvalue.beta=pnorm(beta/sd.beta)
list(beta=beta,sd.beta=sd.beta,pvalue=pvalue.beta)
}
#########################################################################################
# To create distinct Event time for stopped Cox Model and Max Permutation test
##########################################################################################
min.pv.calculation<-function(data, p=seq(0.3,1,by=0.1)){
t0=quantile(unique(data$time[data$event==1]), prob=p)
K=length(t0)
pv=numeric()
for(j in 1:K) pv[j]=stopped.cox(data,t0=t0[j])$pvalue
list(pv=min(pv),t0=t0[which.min(pv)])
}
###########################################################################################
# To Simulate and Replicate the Data
###########################################################################################
BB=1000 # Number of data replication
results=array(, dim=c(BB,10,5)) # To store one sided P-values of all test statistics
Betasss=array(, dim=c(BB,6,5))#To store treatment effects from different models
for(conf in 0:4) # Number of scenarios
for(bb in 1:BB){
```

```
print(bb)
data=simulate.data(n=200,prop.obs=0.7,config=conf, beta.z=0.5)  #If beta.z=0 there are no covariates!
###########################################################################################
#One-sided Overall Treatment and Covariate Effect from Cox PH model
###########################################################################################
fit.overall=coxph(Surv(time, event)~G+z,data)
beta=coefficients(summary(fit.overall))[1,1]
sd.beta=coefficients(summary(fit.overall))[1,3]
#pvalue.beta=coefficients(summary(fit.overall))[5]
pvalue.beta=pnorm(beta/sd.beta)
results[bb,1,conf+1]=pvalue.beta
Betasss[bb,1,conf+1]=beta
#############################################################################################
# EARLY: one-sided Early treatment effect from Stopped Cox at median of event times
#############################################################################################
t0=median(data$time[data$event==1])
ss=stopped.cox(data,t0=t0)
beta.t0=ss$beta
sd.beta.t0=ss$sd.beta
pvalue.beta.t0=ss$pvalue
############################################################################################
# One sided Test Statistic based on Sum of Early and overall treatment Effect
############################################################################################
Z.sum=(beta+beta.t0)/sqrt((sd.beta.t0^2)+3*(sd.beta^2))
results[bb,2,conf+1]=pnorm(Z.sum)
############################################################################################
# EARLY+LATE: Adding a time-dependent treatment in a Cox PH model by heavside function
############################################################################################
t0=median(data$time[data$event==1])
data2=survSplit(data,cut=t0,end="time",event="event",start="start")
data2$gt=(data2$start==t0)+0    #create Heaviside function
fit.td<-coxph(Surv(start,time,event)~G:I(1-gt)+G:gt+z,data=data2)
b=coefficients(summary(fit.td))[-1,1]
s=coefficients(summary(fit.td))[-1,3]
z=coefficients(summary(fit.td))[-1,4]
pv=pnorm(z)
beta.t0=coefficients(summary(fit.td))[2,1]
beta.t0LT=coefficients(summary(fit.td))[3,1]
beta.LT=beta+beta.t0LT
beta.ET=beta+beta.t0
sd.beta.t0LT=coefficients(summary(fit.td))[3,3]
Betasss[bb,2,conf+1]=beta.t0
Betasss[bb,3,conf+1]=beta.t0LT
Betasss[bb,4,conf+1]=beta.ET
Betasss[bb,5,conf+1]=beta.LT
#########################################################################################
# One sided Test Statistic based on Sum of late and overall treatment Effect
#########################################################################################
Z.sumLT=(beta+beta.t0LT)/sqrt((sd.beta.t0LT^2)+3*(sd.beta^2))
results[bb,3,conf+1]=pnorm(Z.sumLT)
##########################################################################################
#a) EARLY: one-sided Early treatment effect from Extended Cox/Stopped Cox at median of event times
##########################################################################################
results[bb,4,conf+1]=pv[1]
###########################################################################################
###b) LATE: one-sided treatment effect from extended Cox/ Left truncated Cox
###########################################################################################
results[bb,5,conf+1]=pv[2]
##########################################################################################
#c) EARLY+LATE: Fisher method of Combining one sided P-values from Extended Cox
```

```
########################################################################################
Fisher.test <- function(p) {
  Xsq <- -2*sum(log(p))
  p.val <- 1-pchisq(Xsq, df = 2*length(p))
  return(c(Xsq = Xsq, p.value = p.val))
}
results[bb,6,conf+1]=Fisher.test(p = pv)[2]
##########################################################################################
# EARLY+OVERALL: One Sided Group sequential design based on  Pvalues from Cox PH and Stopped Cox
#########################################################################################
tau=(sd.beta^2)/(s[1]^2)
alpha=0.05
alpha1=0.03
alpha2=find.alpha2.twosided(alpha1=alpha1,alpha=alpha,tau=tau)
results[bb,7,conf+1]=ifelse(pvalue.beta<alpha1 | pv[1]<alpha2,0,1)
##########################################################################################
# LATE+OVERALL: One Sided group sequential design based on P-values from Cox PH and extended Cox
#########################################################################################
tau=(sd.beta^2)/(s[2]^2)
alpha=0.05
alpha1=0.03
alpha2=find.alpha2.twosided(alpha1=alpha1,alpha=alpha,tau=tau)
results[bb,8,conf+1]=ifelse(pvalue.beta<alpha1 | pv[2]<alpha2,0,1)
##########################################################################################
#One Sided: Maximum of Test statistics from distinct 10 event time points based on Permutation test
##########################################################################################
Cox.TD<-function(data, prob=seq(0.3,1, length=10)){
seq.t0<-as.numeric(quantile(data$time[data$event==1], prob))
pval=numeric()
for(k in 1:length(seq.t0)){
data2=survSplit(data,cut=seq.t0[k],end="time",event="event",start="start")
data2$gt=(data2$start==seq.t0[k])+0    #create Heaviside function
fit.td<-coxph(Surv(start,time,event)~G:I(1-gt)+G:gt+z,data=data2)
b=coefficients(summary(fit.td))[-1,1]
s=coefficients(summary(fit.td))[-1,3]
z=coefficients(summary(fit.td))[-1,4]
pv=pnorm(z)
pv[is.na(pv)]=0.5  #overall Cox: prob=1
pval[k]=Fisher.test(p = pv)[2]
}
opt.pval=min(pval)
}
pv.obs=Cox.TD(data)
PP=300
pv.perm=numeric()
data.perm=data
for(pp in 1:PP){
data.perm$G=sample(data$G)
pv.perm[pp]=Cox.TD(data.perm)
}
results[bb,9,conf+1]=mean(pv.perm<pv.obs)
print(results[bb,,conf+1])
##########################################################################################
# Pseudo values Regression Approach based Restricted mean Survival time
##########################################################################################
library("KMsurv")
library("pseudo")
#########################################################################################
#  To calculate pseudo-values based  mean restricted survival times at 80 percentiles of observed event time
#########################################################################################
```

```
rms.time=quantile(data$time[data$event==1],0.8)
a <- cbind(data,pseudo=pseudomean(data$time, data$event,tmax=rms.time),id=1:nrow(data))
library("geepack")
fit22<- geese(pseudo ~ data$G+data$z, data = a, id=id, jack =T, family=gaussian, corstr="independence",
scale.fix=F)
beta=as.numeric(summary(fit22)[[1]][2,1])
std.beta=as.numeric(summary(fit22)[[1]][2,3])
Zpv = beta/std.beta
results[bb,10,conf+1]=1-pnorm(Zpv)
Betasss[bb,6,conf+1]=beta
}
colnames(results)<-
c('Cox','Z.sum','Z.sumlate','ET','LT','ET+LT(TDTRT)','ET+OT(GS)','LT+OT(GS)','Max.t0','PRMS')
colnames(Betasss)<-c('CoxBeta','Beta.t0','Beta.t0LT','Beta.E+T','Beta.L+T','PRMSBeta')
############################################################################################
#  Summaries of one sided p-values from Simulations
############################################################################################
alpha=0.05
rr=rbind(apply(results[,,1]<alpha,2,mean,na.rm=TRUE),
apply(results[,,2]<alpha,2,mean,na.rm=TRUE),
apply(results[,,3]<alpha,2,mean,na.rm=TRUE),
apply(results[,,4]<alpha,2,mean,na.rm=TRUE),
apply(results[,,5]<alpha,2,mean,na.rm=TRUE))
round(rr,3)
############################################################################################
# Estimated average treatment effects from Simulations
############################################################################################
Trteft=rbind(apply(Betasss[,,1],2,mean,na.rm=TRUE),
apply(Betasss[,,2],2,mean,na.rm=TRUE),
apply(Betasss[,,3],2,mean,na.rm=TRUE),
apply(Betasss[,,4],2,mean,na.rm=TRUE),
apply(Betasss[,,5],2,mean,na.rm=TRUE))
round(Trteft,3)
###########################End###########################################################
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Testing the treatment effect in randomized clinical trials with possible non-proportional hazards

Richting: **Master of Statistics-Biostatistics**
Jaar: **2014**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Anjullo, Belay Belete

Datum: **16/09/2014**