UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

# Faculty of Sciences
## *School for Information Technology*

Master of Statistics

***Masterthesis***

***Statistical inference using generalized pairwise comparisons***

**Abigirl Machingura**

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

**SUPERVISOR :**
Prof. dr. Geert MOLENBERGHS

**SUPERVISOR :**
Prof. Dr. Marc BUYSE

2017
2018

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

**www.uhasselt.be**
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

# Faculty of Sciences
## *School for Information Technology*

Master of Statistics

***Masterthesis***

***Statistical inference using generalized pairwise comparisons***

**Abigirl Machingura**

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

**SUPERVISOR :**

Prof. dr. Geert MOLENBERGHS

**SUPERVISOR :**

Prof. Dr. Marc BUYSE

# Contents

# Acknowledgements

It was not going to be possible to produce this thesis report without the help of the good people that were around me during my research and l wish to extend my gratitudes to them.

It is a great pleasure to express my deep appreciation to my research supervisors **Prof. Marc Buyse**: *International Drug Development Institute (IDDI), San Francisco, CA, USA and Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-Biostat), Hasselt University, Hasselt, Belgium* and **Prof. Geert Molenberghs**: *Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-Biostat), Hasselt University, Hasselt, Belgium* for their guidance, support and encouragement during my research period. I learnt how to research intensively and put all the ideas together into something meaningful. It was not easy but with your help and guidance l was able to meet the objectives that were given to me. I would like to also thank **Prof Tomasz Burzykowski**: *Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-Biostat), Hasselt University, Hasselt, Belgium* for seeing the capacity in me to work on this research.

I would like to thank Professor Marc Buyse for suggesting such a wonderful topic. It was something very new to me and l enjoyed learning new things for my thesis. Thank you for teaching me to understand the topic and for encouraging me to keep on working hard. I would like to also express my gratitude to **Prof Julien Peron**: *MD, PhD Hospices Civils de Lyon, Oncology and Biostatistics departments, Pierre-Benite, France, and University of Lyon 1, CNRS UMR 5558, Biometry and Evolutive Biology Laboratory, Biostatistics-Health Team, Villeurbanne, France*, for allowing us to extend his work on the comparison between generalized pairwise comparisons and the standard log-rank test to comparison between extended generalized pairwise comparisons and weighted log-rank tests. I would like to thank him also for programming assistance to be able to meet the research objectives.

Lastly l would like to extend my appreciation to my family, friends and more importantly my husband Fin for overall support during my research.

**Abstract**

In clinical trials, survival endpoints are often used to assess the effect of new therapies. Several methods have been proposed to compare two or more treatments in the analysis of survival data. Log-rank test have been used traditionally and is still used to compare survival curves under proportional hazards scenario. With an increase in drug development, immunotherapies and vaccines were developed to treat and prolong the lives of cancer patients and other diseases. Such treatments and vaccines mostly take time to have an effect or start working positively, thus they show their effect late during follow up period in a trial. In these situations, the assumption of proportional hazards is violated. Generalized pairwise comparisons and weighted log-rank tests can be used to compare survival curves even when hazards are not proportional over time. The power of the generalized pairwise comparisons was compared to the power of weighted log-rank tests using simulations for three scenarios of survival differences. These scenarios includes proportional hazards, delayed treatment effect and cure rate. The overall hazard ratio for all scenarios was kept the same for these three scenarios for comparability. Under proportional hazards scenario, the standard log-rank test is more powerful than all other tests. The log-rank test loses some power in the presence of censored observations. The net benefit is less powerful than the log-rank test and loses power with increase in threshold of clinical relevance. When there is a delay in treatment effect, the net benefit gains power with an increase in clinical relevance. It is equally powerful to the Fleming and Harrington weighted log-rank test which gives more power to late failure times, for large thresholds of clinical relevance. If a proportion of patients is cured, the net benefit is more powerful than any other test if large thresholds are used. Time of analysis plays an important role in the power of both generalized pairwise comparisons and weighted log-tank tests.

**Keywords: Generalized pairwise comparisons; Net benefit; Weighted log-rank test; Power; Proportional hazards**

# 1 Introduction

The main interest in the analysis of survival data is in the comparison of the patients in two different treatment groups. The main idea is to test whether patients in the new treatment group tend to survive longer or benefit more than patients in the control group. Analysis of clinical trials data is done at a pre-specified time. However some patients experience an event during trial period and some does not. Some patients are lost to follow up and some experience events that makes it difficult to observe an event of interest. This results in censored observations. Estimation and comparison of survival curves is straight forward if all observations are observed completely. However in the presence of censored observations, specialized methods which takes censoring into account are needed in the comparison of survival curves.

The log-rank test is a commonly used method for comparing survival curves and observing treatment benefits when hazards are proportional in time to event endpoints. The method was proposed by Mantel and Haenszel in 1959 and can be used when observations are censored. If there is no censoring, standard two sample tests can be used to compare two treatment groups. The log-rank test is known to be powerful and optimum when hazard ratios are not varying with time. However it loses power when the proportional hazards assumption is not met. When treatment effect is delayed, the proportional hazards assumption will nolonger hold. With an improvement in the drug development, immunotherapies and cancer vaccines for treatment of cancers lead to delays in the treatment effect and therefore leading to violation of the proportional hazards assumption. The overall hazard ratio ceases to have a simple interpretation since it is a function of time. Therefore the log-rank test lose power to detect true treatment benefits.

There are quite a number of scenarios where the non-proportionality of hazards situation is serious. There is a situation where a treatment shows early treatment benefits and loses its effect with time leading to the convergence or crossing of hazards (Yang, S. & Prentice, R., 2010). This leads to violation of the proportional hazards assumption. Another example of such situations is where a treatment takes time to show its effect, that is delayed treatment effect. Thus there is small or no treatment benefit earlier during the trial. Survival curves start to diverge later during the trial follow up for example when a treatment has high deaths due to complications or adverse effects then shows treatments benefits later during follow up. These situations cannot be handled properly by log-rank test since the hazards are no longer proportional over time.

However, researchers came up with various approaches to handle situations where hazards are not proportional over time. These methods includes the weighted log-rank tests and generalized pairwise comparisons(GPC). This study aims at comparing the power of the net benefit which is based on generalized pairwise comparisons to the Fleming and Harrington weighted log-rank tests which is an extension of log-rank test by using different weights on survival times. Both weighted log-rank tests and GPC can

be applied when the proportional hazards assumption is not met. In this study, the power of these tests is assessed using simulated datasets from a randomized clinical trial under three different scenarios of survival differences. These include proportional hazards, delayed treatment effect and cure rate.

In previous studies, it was shown that the GPC has high power compared to the most common log-rank test when there is delayed treatment effect. Weighted log-rank tests which uses different weighting on different failure times was shown to be more powerful than the common log-rank test when treatment effect is delayed(Su, Z. & Zhu, M., 2017). These methods can be applied in medical field for example in cancer screening trials. The UKCTOCS did not manage to reach statistical significance for the primary endpoint with the log-rank test in the screening trial for ovarian cancer because late screening effects were not taken into account(Jacobs, I J., 2015). The log-rank test did not have enough power to detect delayed treatment effects. Thus using methods which takes delayed treatment effects into account will give valid and more relevant estimates of the treatment effects.

Generalized pairwise comparisons and weighted log-rank tests are described in detail in section two showing how each method works together with advantages and limitations. Comparison of generalized pairwise comparisons and weighted log-rank test by simulations is described in section three for each scenario of survival difference. Results interpretation and discussion is described in section four. The final section gives the conclusive discussion and recommendations with respect to future similar studies.

# 2 Methodology

## 2.1 Generalized pairwise comparisons

Generalized pairwise comparisons extends non-parametric tests and lead to a general measure of the difference between the treatment groups called the "proportion in favor of treatment" or "net benefit" which is related to traditional measures of treatment effect for a single variable (Buyse, M., 2010). The method extends the U-statistics of the Wilcoxon Mann Whitney test. The Wilcoxon Mann Whitney can be applied to both continuous and binary outcomes when there is no censoring in the data and cannot be applied if there are censored observations in the data since it would be not possible to rank observations. Some observations might not have experienced an event during the time of comparison. One way to handle such situations, suggested by Gehan(1965), is to use pairwise comparisons. Thus with generalized pairwise comparisons, data with censored observations can be analyzed and also allowing group comparisons for variables of any type.

In generalized pairwise comparisons, there is comparisons of all possible pairs of individuals with one from treatment group and the other from the control group. A pair of individuals is considered favorable if an individual from the the treatment group is better than an individual from the control group by a certain amount larger than a clinically relevant threshold that is pre-specified. A pair is considered unfavorable if an individual from the control group is better than the one from the treatment group. It is also considered neutral if no difference is observed between the two individuals from the two groups. However, using an extension of generalized pairwise comparisons, a pair of censored observations is not considered uninformative but rather uses the magnitude of the censored observations to calculate the score values. Let $X_i$ and $Y_i$ be observed outcomes of a time to event outcome measure with $X_i$ being the observed outcome from the control group for $i = 1, 2, ..., n$ and $Y_i$ from the treatment group for $j = 1, 2, ..., m$. Thus the pairwise score indicator for each pair of control outcome and treatment outcome is given by

$$U_{ij} = \begin{cases} +1 & \text{if pair } (X_i, Y_j) \text{ is favorable} \\ -1 & \text{if pair } (X_i, Y_j) \text{ is unfavorable} \\ 0 & \text{Neutral} \end{cases}$$

Table 1 shows calculations of the pairwise score values using extended generalized pairwise comparisons proposed by Peron *et al* (2016), which makes better use of censored observation by considering the information available for the censored observations. When a pair has both observations censored, the extended generalized pairwise comparisons makes use of the available information by considering the magnitude of the censored observations. Let $X_i$ and $Y_j$ be uncensored observations and $X_i'$ and $Y_j'$ be censored observations with a pre-specified clinically relevant threshold $\epsilon$.

Table 1: Pairwise score values ($U_{ij}$) computed using the extended procedure for generalized pairwise comparisons of a time to event outcome

| Censoring | Pairwise Comparison | $U_{ij}$ |
|---|---|---|
| No censoring | $X_i - Y_j \geq \epsilon$ | +1 |
| | $X_i - Y_j \leq -\epsilon$ | -1 |
| | $\lvert X_i - Y_j \rvert < \epsilon$ | 0 |
| X censored and Y not censored | $X_i' - Y_j \geq \epsilon$ | +1 |
| | $X_i' - Y_j \leq -\epsilon$ | $\frac{\hat{S}_T(y_j+\epsilon)+\hat{S}_T(y_j-\epsilon)}{\hat{S}_T(X_i)} - 1$ |
| | $\lvert X_i' - Y_j \rvert < \epsilon$ | $\frac{\hat{S}_T(y_j+\epsilon)}{\hat{S}_T(X_i)}$ |
| Y censored and X not censored | $X_i - Y_j' \geq \epsilon$ | $1 - \frac{\hat{S}_C(x_i+\epsilon)+\hat{S}_C(x_i-\epsilon)}{\hat{S}_C(y_j)}$ |
| | $X_i - Y_j' \leq -\epsilon$ | -1 |
| | $\lvert X_i - Y_j' \rvert < \epsilon$ | $-\frac{\hat{S}_C(x_i+\epsilon)}{\hat{S}_C(y_j)}$ |
| Both X and Y Censored | $X_i' - Y_j' \geq \epsilon$ | $1 - \frac{\hat{S}_C(x_i-\epsilon)}{\hat{S}_C(y_j)} - \int_{t>x_i-\epsilon}^{\infty} \frac{\hat{S}_T(t+\epsilon)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t) + \int_{t>x_i}^{\infty} \frac{\hat{S}_C(t+\epsilon)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_T(t)$ |
| | $X_i' - Y_j' \leq -\epsilon$ | $-\int_{t>y_i}^{\infty} \frac{\hat{S}_T(t+\epsilon)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t) + \frac{\hat{S}_T(y_j-\epsilon)}{\hat{S}_T(x_i)} + \int_{t>y_j-\epsilon}^{\infty} \frac{\hat{S}_C(t+\epsilon)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_T(t) - 1$ |
| | $\lvert X_i' - Y_j' \rvert < \epsilon$ | $\int_{t>y_j}^{\infty} \frac{\hat{S}_T(t+\epsilon)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_C(t) + \int_{t>x_i}^{\infty} \frac{\hat{S}_C(t+\epsilon)}{\hat{S}_T(x_i)\hat{S}_C(y_j)} d\hat{S}_T(t)$ |

### 2.1.1 Net benefit

Net benefit is used to measure the treatment benefit and is defined as the probability that a random patient in the treatment group survives longer than a random patient in the control group minus the probability of the opposite situation(Peron, J., 2016). The net benefit is estimated using generalized pairwise comparisons. It shows the net chance of surviving of a patient on treatment than when not on treatment. The net benefit is estimated by

$$\hat{u} = \Delta = \frac{1}{n.m} \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij} \tag{1}$$

which is the difference between the number of favorable and unfavorable pairs divided by the overall number of pairs. For $\Delta$=0, the treatment group and the control group are not different. For $\Delta > 0$, the treatment group is better than the control group and for $\Delta < 0$, the control group is better than the treatment group. If the calculated net benefit is 0.05, it means that a random patient from the treatment group has 5% more chance or probability of surviving longer than a random patient in the control group. The net benefit does not depend on the type of variables considered thus it can be calculated for any type of outcome variable. It is closely related to the probabilistic index which is also a measure of treatment effect. The net benefit is advantageous in that it is easier to interpret than probabilistic index and a net benefit of zero implies no treatment effect.

## 2.2 Weighted log-rank test

A frequently used statistical method for the analysis of clinical data is a non-parametric test to compare the survival times for the two patient groups. The log-rank test is the commonly used test to compare survival curves (Schoenfeld D, 1981). It gives equal weights to all failure times in the trial. It assumes that the hazard functions for the two patient groups are not changing with time. However, when the proportional hazards assumption is not satisfied, log-rank test loses power and the standard

Cox model generally produces biased estimates under such conditions (Lin R S& Leon L F, 2017). Some situations in survival time such as delayed treatment effect violates the proportional hazards assumption.

The log-rank test compares outcomes over the whole time interval and may not adequately detect important differences between groups which occur either early or late in the interval (Karadeniz, P G. & Ercan, I., 2017). However weighted log-rank tests are used in situations where the proportional hazards assumption does not apply by using different weighting on failure times. They increase power and sensitivity when the hazard functions depends on time. They also allow for the inflation of early or late treatment differences at an optimal power. Weighted log-rank tests preserves type 1 error when the hazards are not proportional. However the method can be sub-optimal when the weights are not specified correctly. The choice of weights for the log-rank tests is far from intuitive.

Table 2: Summary of events at time $t_k$

| Treatment Group | Treatment | Control | Total |
|---|---|---|---|
| Number of events | $d_{1k}$ | $d_{2k}$ | $d_k$ |
| Number at risk | $r_{1k}$ | $r_{2k}$ | $r_k$ |

Table 2 gives summary of events which are calculated repeated for all time points. $d_{1k}$ and $d_{2k}$ are the number of individuals who experienced an event of interest in treatment and control group respectively at time $k$. $r_{1k}$ and $r_{2k}$ are the number of individuals at risk in the treatment and control group respectively at time $k$. $d_k$ and $r_k$ are the number of individuals who experienced an event of interest and number of individuals at risk respectively in both groups at time $k$. Expected and observed events are also calculated from this table for each treatment.

$$O_{ik} = d_{ik} \tag{2}$$

$$E_{ik} = d_k \frac{r_{ik}}{r_k} \tag{3}$$

$O_{ik}$ is the number of observed outcomes in group $i$ at time $k$, $E_{ik}$ is the number of expected outcomes in group $i$ at time $k$, $W_k$ are weights at time $k$, observed from time $t_1 < t_2 < ... < t_s$. The test statistic for the weighted log-rank test is given by

$$\frac{[\sum_{k=1}^{s} W_k(O_{ik} - E_{ik})]^2}{\sum_{k=1}^{s} var(W_k(O_{ik} - E_{ik}))} \tag{4}$$

$W_k$ are weights at time $k$.

Different weighting choices in the weighted equations for different tests have been proposed. The log-rank test uses equal weighting for both early and late failure times. The Gehan Generalized Wilcoxon test proposed by Gehan in 1965 uses the number of patients at risk as weights. Tarone-Ware in 1977 also proposed a test which uses the number of patients at risk as weights by taking the square root of the number of patients at risk. The weights given by Tarone-Ware test are smaller than those given by Gehan

Generalized Wilcoxon test. The Peto-Peto test uses the modified Kaplan Meier estimator as weights. The Fleming and Harrington class of weights uses the Kaplan Meier estimate raised to a specified power to calculate the weights. Details of each of the tests and test statistics are given below.

- **Log-rank** (Mantel, N.& Haenszel, W.,1959)

  The Log-rank test is used to compare survival curves when hazards are proportional over time. It assumes that the hazard functions are parallel hence hazard ratios for the two treatment groups are constant for all the time points. The weight function is given by $W_k = 1$. Thus there is equal weighting of the earlier failure times and late failure times since the hazards are constant over time. The test statistics for the log-rank test is given by

$$\frac{[\sum_{k=1}^{s}(O_{ik} - E_{ik})]^2}{\sum_{k=1}^{s} var((O_{ik} - E_{ik}))} \tag{5}$$

  There is no weight function in the test statistics equation because there is equal weighting of the failure times.

- **Gehan Generalized Wilcoxon Test**(Gehan A, 1965)

  Is a distribution-free two-sample test which is an extension of the Wilcoxon test to samples with arbitrary censoring on the right (Gehan A, 1965). It uses total number of patients at risk at each failure time as weights, that is $W_k = r_k$. The Gehan Generalized Wilcoxon Test put more emphasis on early survival times where the number of people at risk is larger since it uses the number at risk as weights. Thus more weight is given to early survival times than late survival times where the number of individual at risk is small.

$$\frac{[\sum_{k=1}^{s} r_k(O_{ik} - E_{ik})]^2}{\sum_{k=1}^{s} var(r_k(O_{ik} - E_{ik}))} \tag{6}$$

  where $r_k$ is the total number of patients at risk at time $k$.

- **Tarone-Ware** (Tarone R E& Ware J, 1977)

  The Taron-Ware tests also uses the number of individual patients at risk as weights. The weights are given as the square root of the number of individuals at risk, that is $W_k = \sqrt{r_k}$. More weight is given on earlier survival times than late since the number of individuals at risk are more at the earlier than later. The weights used are smaller than those used by the Gehan Generalized Wilcoxon Test.

$$\frac{[\sum_{k=1}^{s} \sqrt{r_k}(O_{ik} - E_{ik})]^2}{\sum_{k=1}^{s} var(\sqrt{r_k}(O_{ik} - E_{ik}))} \tag{7}$$

- **Peto-Peto** (Peto, R., & Peto, J., 1972)

  The Peto-Peto test is used when hazards are not proportional and uses the estimation of survival function, the modified version of the Kaplan-Meir estimator, as weights (Karadeniz, P G. & Ercan, I., 2017). The weight function is given by $W_k = S(\hat{t}_k)$. Thus earlier failure times receives larger

weights than late failure times because survival functions are larger at the earlier.

$$\frac{[\sum_{k=1}^{s} S(\hat{t}_k)(O_{ik} - E_{ik})]^2}{\sum_{k=1}^{s} var(S(\hat{t}_k)(O_{ik} - E_{ik}))} \tag{8}$$

where $S(\hat{t}_k)$ is the modified Kaplan Meier estimator at time $k$.

- **Fleming & Harrington family,** $G^{\rho,\gamma}$(Fleming T R & Harrington D P, 1991)
  The $G^{\rho,\gamma}$ uses $W_k = (\hat{S}(t_{k-1}))^{\rho}(1 - \hat{S}(t_{k-1}))^{\gamma}$ as weight function and has two parameters $\rho \geq 0$ and $\gamma \geq 0$. For $\rho = \gamma = 0$, then we have the standard log-rank test, for $\rho = 1$ and $\gamma = 0$ then we have Gehan's Generalized Wilcoxon test.

Some treatments reduce the hazard function in the earlier periods of the follow-up and the treatment effect becomes negligible later during follow-up (Karadeniz, P G. & Ercan, I., 2017). Thus methods such as Generalized Wilcoxon Test and Tarone-Ware test which gives more weight to earlier failure times can be used when earlier treatment effect is expected. The $G^{\rho,\gamma}$ is flexible and it can be adjusted for either early, middle or late survival times to have more weights. This study focuses on $G^{\rho,\gamma}$ class of weights proposed by Fleming and Harrington (1981). The $G^{\rho,\gamma}$ class of weights is much more flexible in detecting early and late survival times. For large values of $\rho$, early differences in survival differences can be detected. For large values of $\gamma$, late differences in survival differences can be detected. For equal values of $\rho$ and $\gamma$, more weights are given to failure times occurring at the middle of the overall follow up time.

## 2.3 Fleming and Harrington family($G^{\rho,\gamma}$)

The Flemming and Harrington family of weights is a subclass of weighted log-rank statistics proposed by Fleming and Harrington (1981) used to compare survival curves between two treatment groups. The $G^{\rho,\gamma}$ class of weights is much flexible in detecting early and late survival differences and is less powerful than the standard log-rank test when hazards are proportional over time. The weights should be specified prior to collection of the data for the results to be meaningful. In Fine *et al* 2006, it was shown that weighted log-rank tests under $G^{\rho,\gamma}$ are more powerful than standard log-rank test when treatment effect is delayed even by any small delay. Choice of weights for the parameters $\rho$ and $\gamma$ are illustrated in Figure 1. Higher values of $\rho$ results in detection of early differences in survival times and higher values of $\gamma$ results in detection of late differences in survival times.

For the $G^{\rho,\gamma}$ class of weights, the weight function is chosen to be the Kaplan-Meier estimate of the survival function raised to a specified power (Buyske S, Fagerstrom R & Ying Z, 2012). The standard log-rank test($\rho = 0, \gamma = 0$) and the Peto Prentice ($\rho = 1, \gamma = 0$) are special cases of the $G^{\rho,\gamma}$ class of weights. It is more advantageous because choosing or changing weights leads to increased efficiency than other tests because the choice of $\rho$ and $\gamma$ determines where much or less weight is going.

The weights for the $G^{\rho,\gamma}$ class of weights is given by

$$W_k = (\hat{S}(t_{k-1}))^{\rho}(1-\hat{S}(t_{k-1}))^{\gamma} \tag{9}$$

At time $k$, the survival function is evaluated at a previous time to failure $t_{k-1}$. The test statistic for the $G^{\rho,\gamma}$ is given by

$$\frac{[\sum_{k=1}^{s}(\hat{S}(t_{k-1}))^{\rho}(1-\hat{S}(t_{k-1}))^{\gamma}(O_{ik}-E_{ik})]^2}{\sum_{k=1}^{s} var((\hat{S}(t_{k-1}))^{\rho}(1-\hat{S}(t_{k-1}))^{\gamma}(O_{ik}-E_{ik}))} \tag{10}$$

where

$$\hat{S(x)} = \prod_{t_k \leq t}\left(1-\frac{d_k}{r_k}\right) \tag{11}$$

is the estimator of the Kaplan-Meier survivor function.

Figure 1 shows different weight choices for $\rho$ and $\gamma$. Thus the description of each of the scenarios is given below.

(a) $\boldsymbol{\rho} = 0$ and $\boldsymbol{\gamma} = 0$: There is equal weighting of all the failure times, thus early and late survival times are weighted equally. It is equivalent to the standard log-rank test. The weight function is given by

$$W_k = 1 \tag{12}$$

(b) $\boldsymbol{\rho} = 1$ and $\boldsymbol{\gamma} = 0$: More weights are given to earlier failure times. Failure times that happens late are given small weights because larger hazard functions are found at the beginning. It is equivalent to the Peto-Peto test. The weight function is given by

$$W_k = \hat{S(t)} \tag{13}$$

(c) $\boldsymbol{\rho} = 1$ and $\boldsymbol{\gamma} = 1$: . More weight is given to failure times that happens at the middle of the total follow up time. Failure times that happens early and late are given smaller weights than the ones at the middle. The weight function is given by

$$W_k = \hat{S(t)}(1-\hat{S(t)}) \tag{14}$$

(d) $\boldsymbol{\rho} = 0$ and $\boldsymbol{\gamma} = 1$: More weights are given to late failure times and less weights are given to earlier failure times. The weight function is given by

$$W_k = (1-\hat{S(t)}) \tag{15}$$

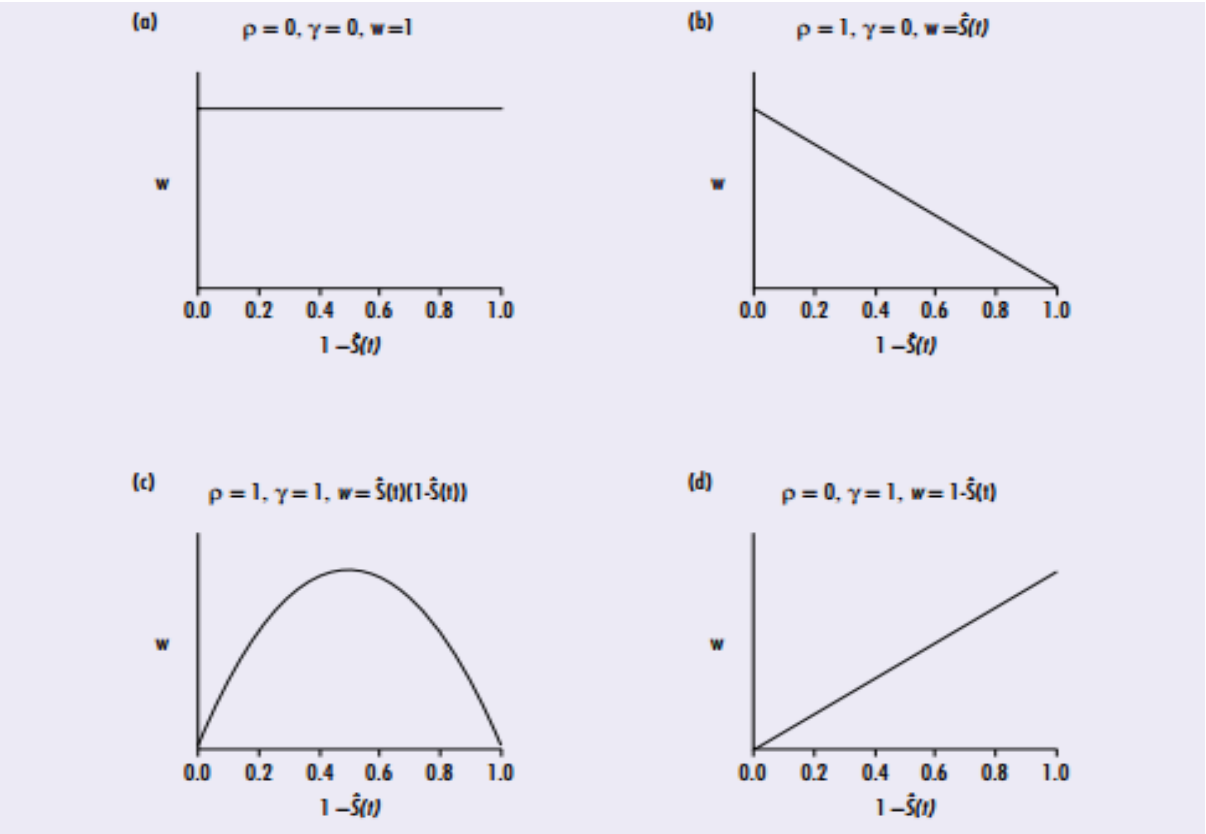


Figure 1: Fleming and Harrington family choice of weights

# 3 Comparison of methods by simulations

In an attempt to find the most powerful test when there are scenarios of survival differences, 1000 datasets of each scenario were generated by simulation. The scenarios include the proportional hazards, delayed treatment effect and cure rate scenario. Each of the 1000 datasets includes two treatment groups, the control group and the treatment group, each with 100 patients. Administrative censoring was introduced by arbitrarily choosing two dates of analysis so that there is 0% and 20% censoring in the dataset. The initial hazard ratio was arbitrarily chosen to be 0.65. Thus for all the three scenarios, the proportional hazards parameters are chosen in such a way that the overall hazard ratio remains at 0.65 so that the scenarios of the survival differences can be comparable. For the first scenario of proportional hazards, the hazard ratio was kept constant at 0.65. The second scenario, there is a delay in the effect of a treatment. Thus the hazards are no-longer proportional. The hazard ratio stayed at 100% for the first 4 months of the follow up time showing no treatment effect and decreased to 30% in later times of the follow up showing delayed treatment effect. The third scenario consists of overall cure of a proportion of patients. The hazards dropped continuously from 100% to exactly 0% in a year of follow up. For each dataset in each scenario, the net benefit and weighted log-rank tests were computed and tested for statistical significance. The common log-rank test which is a special case of the weighted log-rank tests was also included for comparison. The power of the test was calculated by considering the proportion of the tests, in a thousand tests, that have a statistical significance $p$-value less than 0.05 level of significance. The larger the proportion of $p$-values less than 0.05 the smaller the probabilities of false positive that is the probability of making a wrong conclusion that there is difference between treatments when there is no real difference. The power of the net benefit and weighted log-rank tests were plotted under the three scenarios of survival differences. The power was plotted against different thresholds of clinical relevance($t$) in months.

# 4 Results

## 4.1 Survival curves

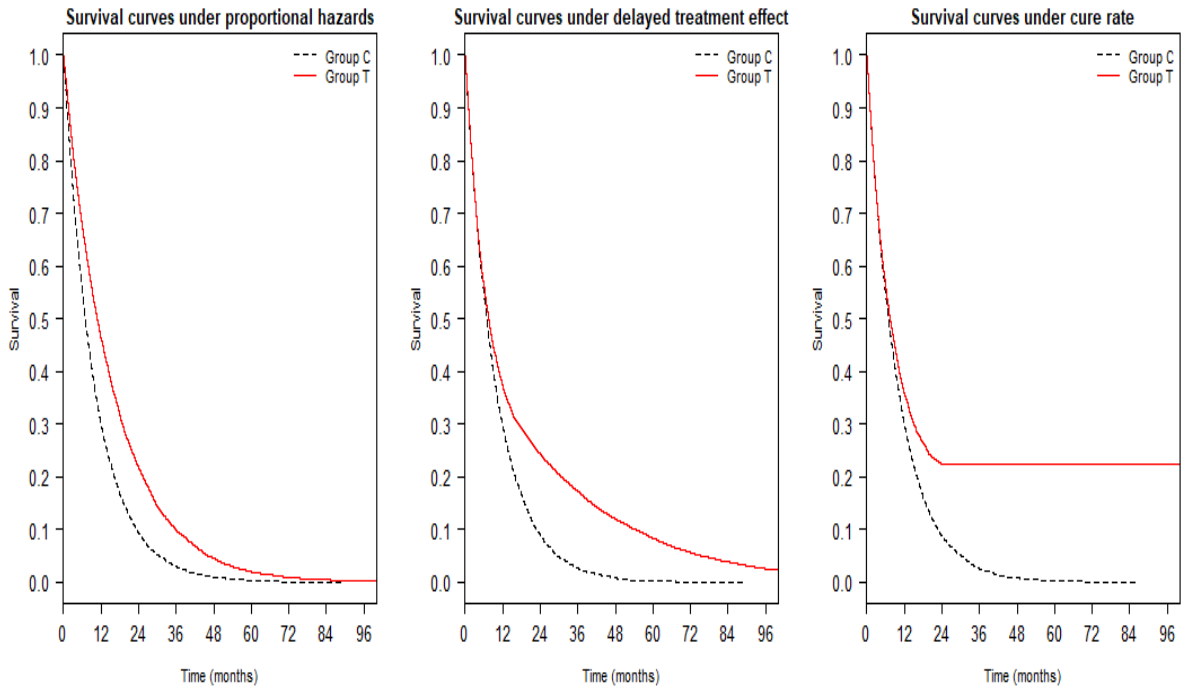


Figure 2: Survival curves for 3 scenarios of survival differences

Figure 2 shows the survival probabilities for the two treatment groups plotted under three scenarios of survival differences that is proportional hazards, delayed treatment effect and overall cure. The first plot shows the survival curves under the proportional hazards. In the plot, the treatment group shows to have survival benefit over the control group. The log-rank test can be used to assess the survival curves under this scenario as hazards have shown to be proportional over time. The middle plot shows survival probabilities of the two treatment groups in the presence of delayed in the treatment effect. The plot shows that there is no treatment difference for the first four months of the trial and curves start to diverge later on after four months of the follow up. The treatment group shows to be much more beneficial than the control group as it is shown by a large divergence of the survival curves. The last plot shows survival probabilities when there is overall cure for a proportion of patients showing a continuous difference between the treatment and the control group with much large treatment benefits from the treatment group for later follow up times.

## 4.2 Estimation of treatment effect

Analysis of survival data was done to estimate treatment effect for the 200 patients in both treatment groups. For all three scenarios of survival difference, patients in treatment group survived longer than patients in control group. This is shown by a significant median survival time of 10.76 months for patients in treatment group compared to 6.96 months for patients in control group under proportional hazards scenario(confidence interval $6.78 - 11.04$ months). The median survival time under the delayed treatment effect is 7.59 months for patients in treatment group compared to 6.95 months for patients in control group. The same patterns are observed for the cure rate scenario where patients in treatment group tends to have longer significant median survival time than the ones in the control group. In figure 2, under delayed treatment effect and cure rate scenario, a proportion of patients have shown to gain late or long term treatment benefits as illustrated by late divergence of survival curves.

Figure 3 shows the overall net benefit for each scenario of survival difference. For proportional hazards, the overall net benefit decreased with an increase the threshold of clinical relevance. When any survival difference is considered clinically relevant(threshold of clinical relevance=0 months), the net benefit is 21%. This means a patient in the treatment group has 21% more chance of surviving longer than a patient in the control group when any survival benefit is considered relevant. However, The overall net benefit drops down to 14% when considering a threshold of clinical relevance of 24 months. This means that the power to detect true treatment benefit decrease with an increase in threshold of clinical relevance when hazards are proportional. The same patterns are also observed in the presence of censored observations were there is a lower treatment benefit than in the absence of censored observations.

In the presence of delayed treatment effect(Figure 3, top right), the net benefit increased with an increase in the threshold of clinical relevance and tend to decrease when long-term thresholds are considered. When any treatment benefit is considered clinically relevant, the net benefit is 12% which means that a patient in the treatment group has 12% chance of surviving longer than a patient in the control group. At a threshold of 24 months, the net benefit is 15%. The net benefit under delayed treatment effect is maintained around 11% when longer thresholds of clinical relevance are considered. In this case, same patterns are observed in the presence of censored observations.

When a proportion of patients is cured, as shown in Figure 3 bottom left, the net benefit increases with an increase in threshold of clinical relevance. For small thresholds of clinical relevance, the net benefit is small and is more pronounced for long-term thresholds of clinical relevance. When any treatment is considered clinically relevant, the net benefit is 11%. When a threshold of 24 months is considered the net benefit is 20% in favor of the treatment group. For long term thresholds of clinical relevance, the net benefit is even more higher. When a threshold of 42 months is considered clinically relevant, the net benefit is 24% implying that a patient in the treatment group benefit much more than a patient in the control group. When some of the observations are censored, the same pattern follows for the cure rate scenario.

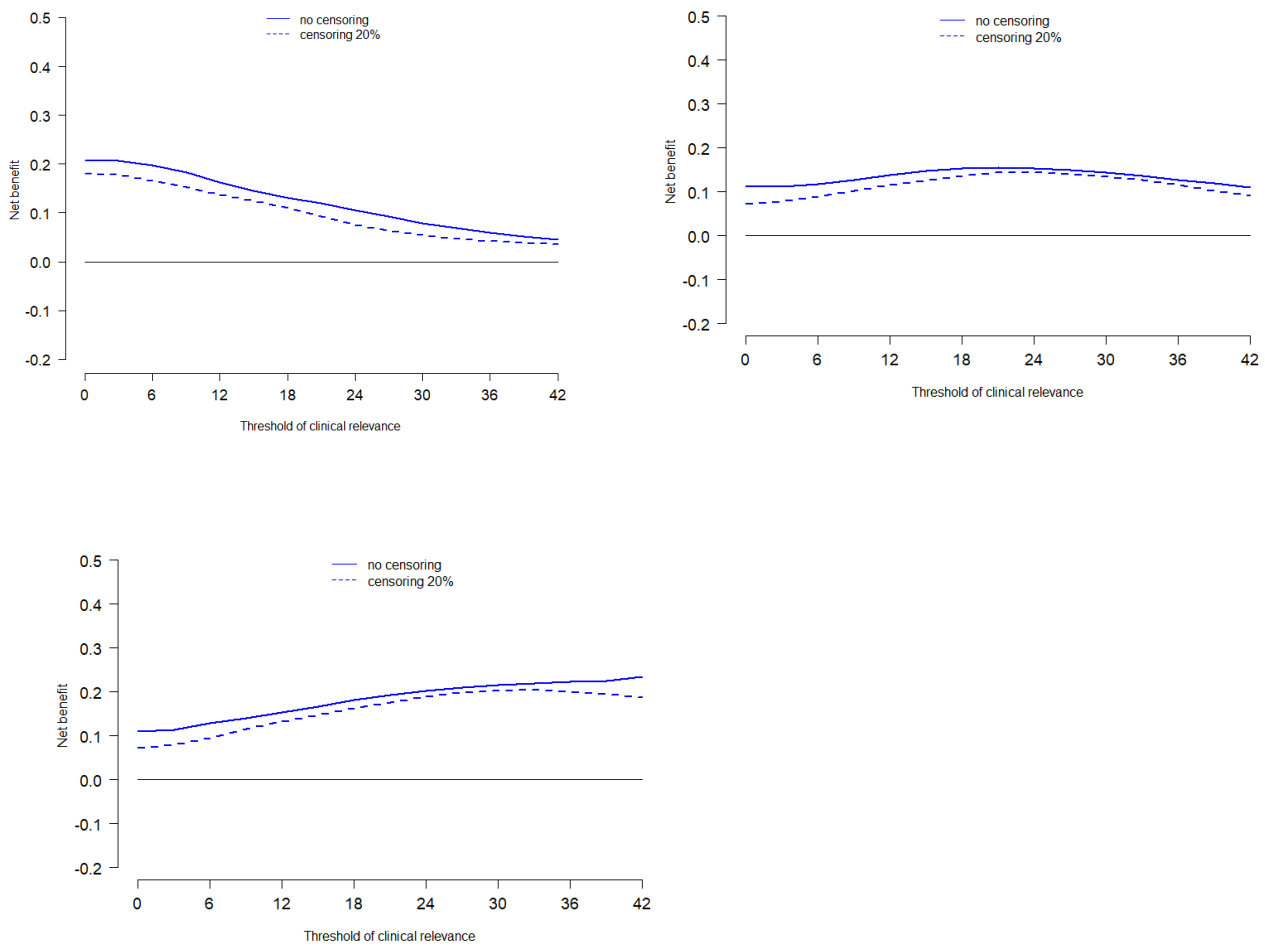


Figure 3: Net survival benefit for each scenario of survival difference: Proportional hazard, top left; delayed treatment effect, top right and cure rate, bottom left.

### 4.3 Scenario 1: Proportional Hazards

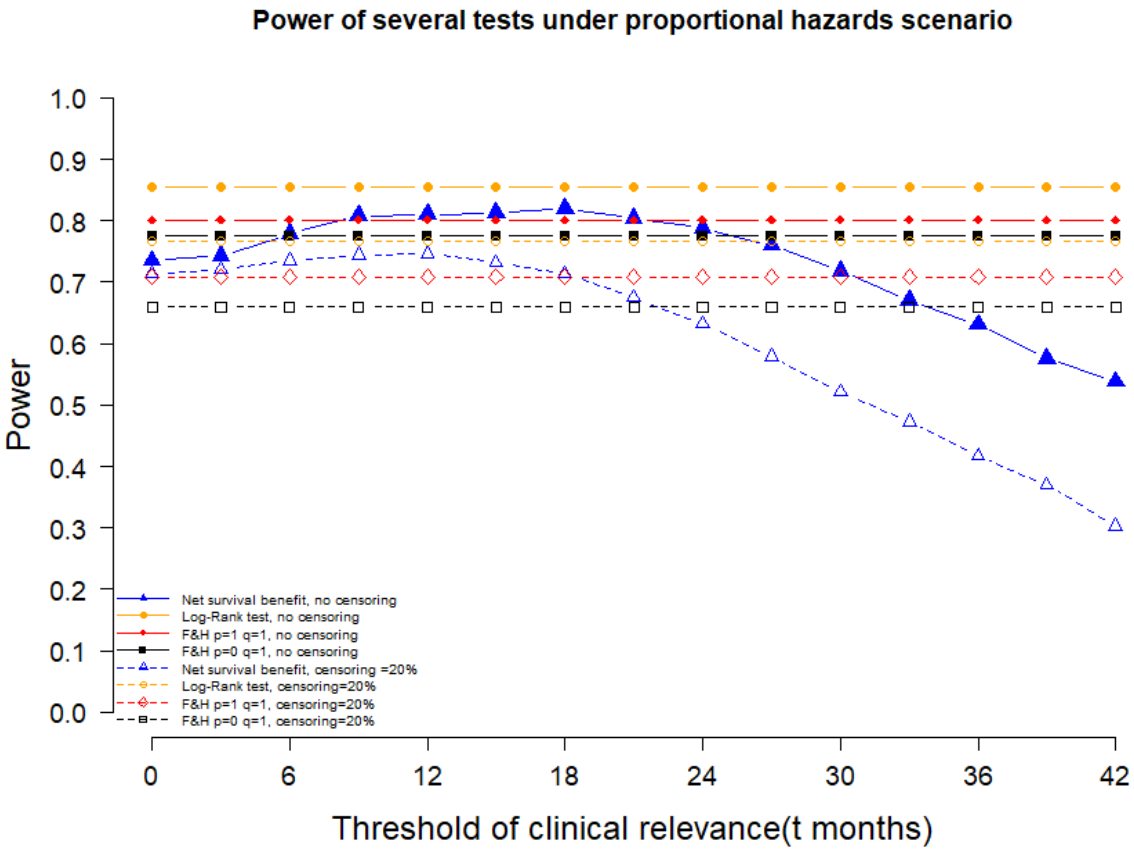


Figure 4: Survival curves under proportional hazards scenario

Survival curves produced under proportional hazards scenario are shown in Figure 4. Four different tests were used to produce and compare the survival curves. The tests includes net benefit estimated by GPC, log-rank test, the Fleming and Harrington $G^{1,1}$ and the Fleming and Harrington $G^{0,1}$. Under proportional hazards, the hazards ratios for the two groups of treatments should be constant thus parallel survival functions. In this case, the log-rank test has the highest power of detecting treatment effect than all other tests with a power of 86% when there is no censoring and 77% in the presence of censored observations. When considering any survival difference, that is $t = 0$ threshold of clinical relevance, the net benefit has a power of 74%. The power increased with an increase in the threshold to 82% at a threshold of 18 months. Thus at a threshold of 18 months, the net benefit is almost as powerful as the standard log-rank test under proportional hazards assumption. When there are censored observations, the log-rank test remains the most powerful followed by the net benefit with power 77% and 75% respectively. From month 18, the net benefit starts to decrease in power with an increase in the threshold of clinical relevance. At month 42 the power for the net benefit dropped to 54%. In this case the Fleming and Harrington $G^{0,1}$ is the least powerful and it is more affected by the presence of censored observations. The power of the weighted log-rank tests remained constant with changes in the threshold of clinical relevance but the net

benefit did depend on the threshold.

### 4.4 Scenario 2: Delayed Treatment effect

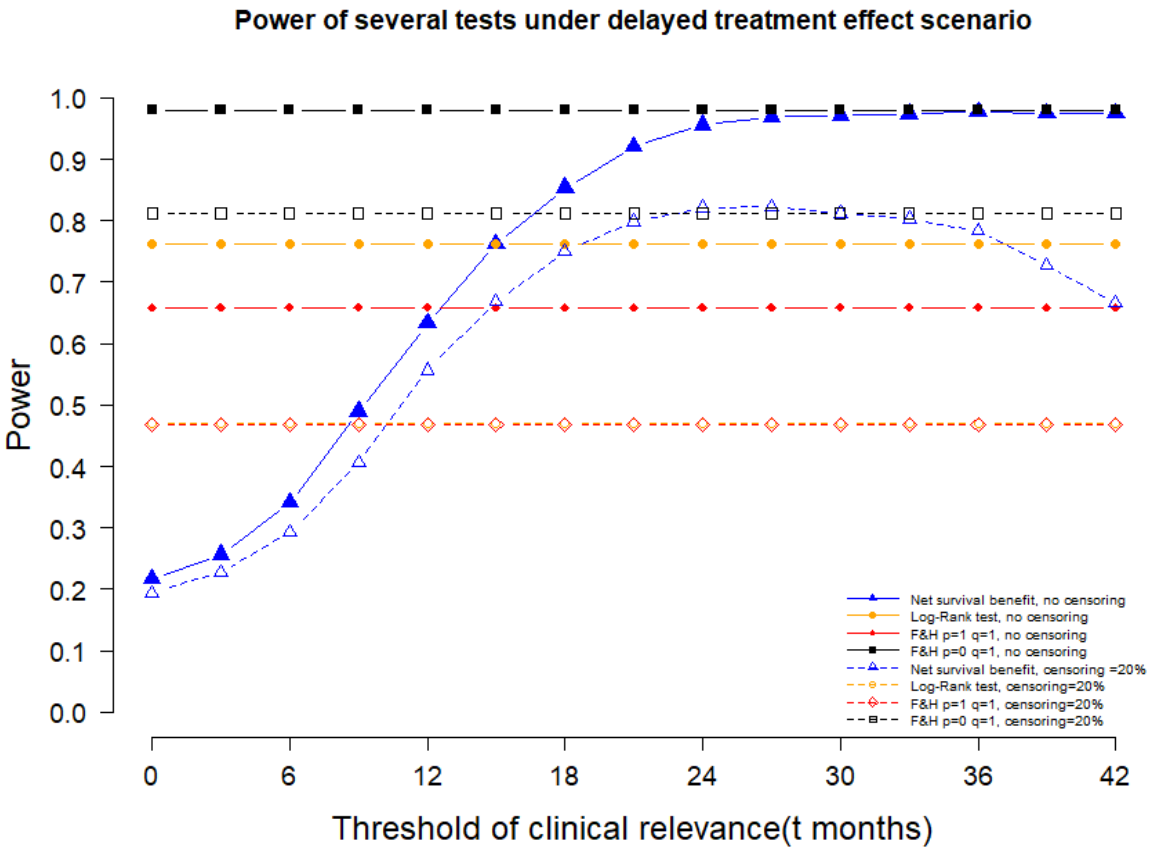


Figure 5: Survival curves under delayed treatment effect scenario

Figure 5 shows survival power curves produced in the presence of delayed treatment effect using net benefit, log-rank test, the Fleming and Harrington $G^{1,1}$ and the Fleming and Harrington $G^{0,1}$. At a threshold of month $t = 0$, the net benefit has the lowest power as compared to all other test with the Fleming and Harrington $G^{0,1}$ having the highest power of 98% as compared to 22% when there no censoring and 81% as compared to 19% when there is censoring. The log-rank test and the Fleming and Harrington $G^{1,1}$ also performed better than net benefit at threshold $t = 0$. The power of the net benefit increased with an increase in the threshold of clinical pertinence. At month $t = 24$, the power of net benefit rose from 22% to 96% in the absence of censoring and from 19% to 82% when there is 20% censoring. It became more powerful than the standard log-rank test and the Fleming and Harrington $G^{1,1}$ except for the Fleming and Harrington $G^{0,1}$ which remained the most powerful than all the tests. At month $t = 42$, the power of the net benefit rose to 98% with no censoring and 67% when there is censoring and starts to decrease for very long thresholds that is $t > 42$ months. The Fleming and Harrington $G^{0,1}$ remained the most powerful even for very long term thresholds. The Fleming and Harrington $G^{0,1}$ lost

power in the presence of censored observations that is from 98% to 81% in the presence of 20% censoring. The standard log-rank test and the Fleming and Harrington $G^{1,1}$ are the least powerful in the presence of delayed treatment effect with power of 76% and 66% when there is no censoring and 47% and 47% in the presence of censored observations respectively. The log-rank test lost more power than all other tests in the presence of censored observations.

### 4.5 Scenario 3: Cure rate

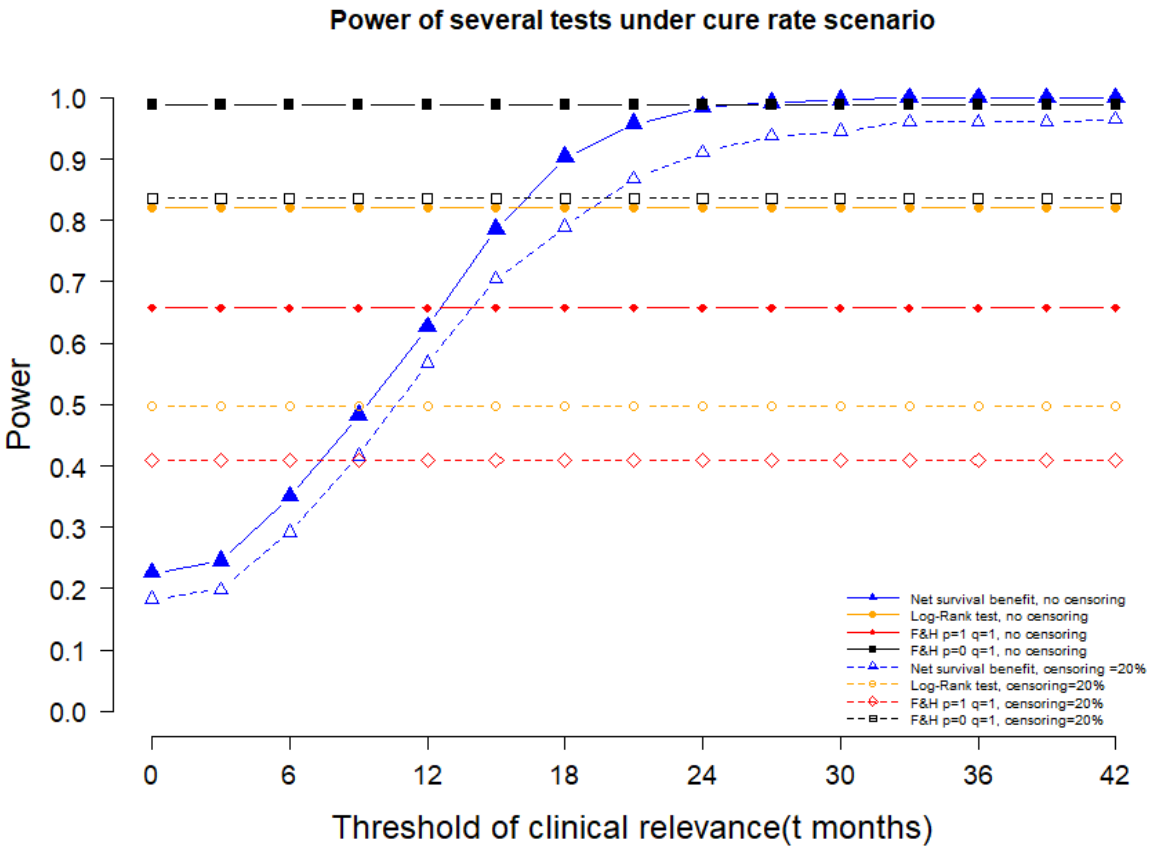


Figure 6: Survival curves under cure rate scenario

Figure 6 shows survival curves plotted under cure rate scenario. The curves are produced from the four tests which includes net benefit, log-rank test, the Fleming and Harrington $G^{1,1}$ and the Fleming and Harrington $G^{0,1}$ in the presence of 0% and 20% administrative censoring. When the threshold of clinical relevance is zero($t = 0$ months), that is any survival benefit is considered clinically relevant. When considering any treatment benefit to be clinically relevant, the net benefit has the lowest power as compared to all other tests with a power of 23% in the absence of censoring and 18% in the presence of censoring. The Fleming and Harrington $G^{0,1}$ has the highest power 99% and 84% when observations are censored at a threshold of 24 months. As the large survival differences are considered relevant, the power of the net benefit started to increase. Thus at a threshold of 24 months($t = 0$ months), the power

of net benefit rose from 23% to 98% in the absence of censored observations and from 18% to 91% in the presence of censored observations. As the threshold of clinical relevance is further increased, the power of net benefit is more evident and noticeable. The power of net benefit rose to 99.9% when a clinical relevant survival difference of 42 months is considered and 96% in the presence of censored observations. The net benefit became more powerful than weighted log-rank tests when large thresholds are considered. The log-rank tests and the Fleming and Harrington $G^{1,1}$ are the least powerful with power 82% and 66% when there are no censored observations and 50% and 41% in the presence of censored observations respectively. The net benefit did not lose much power in the presence of censored observations (99.9% to 96%) compared to any other tests under the cure rate scenario. It is worth noting that in the presence of cure rate, the net benefit tends asymptotically to the cure rate, and as such provides a naturally and intuitively appealing way of testing long-term treatment benefits.

# 5 Discussion

The main aim in the analysis of survival data is to observe if patients on new treatment are deriving benefit as compared to the ones on the control treatment. Log-rank test has been traditionally used to compare two survival curves for the efficacy of the new treatment over the control. However the log-rank test is only optimal if the proportional hazards assumption is not violated. With growth in the drug development for example in immunotherapies and cancer vaccines, the proportional hazards assumption is violated. Vaccines and some drugs takes time to start working or produce positive results in the human system thus leading to delays in the treatment effect. This also happens with treatments that are aggressive at the beginning of consumption by a patient and shows their effect or positive results later during follow up. These situations and many more leads to violation of the proportional hazards assumption hence the log-rank test loses power and results interpretation will be meaningless.

Generalized pairwise comparisons and weighted log-rank tests can be used to compare survival curves even if the proportional hazards assumption is violated. These methods relaxes the assumption of proportional hazards hence more powerful than the standard log-rank test when hazards are not proportional. Generalized pairwise comparisons estimates the net chance of a better outcome (net benefit) with treatment group than with control group by comparing the patients outcomes among all possible pairs taking one patient from the treatment group and one patient from the control group. The net benefit is a measure of treatment effect which is defined as the expected proportion of patients for which the outcome is higher in the treatment group minus the expected proportion of patients for which the outcome is higher in the control group(Peron, J., 2016). The method is used as an alternative to other non-parametric tests to assess treatment effect between two groups in the presence of censored observations. With generalized pairwise comparisons the outcome measure variable of any type for example time to event, binary, continuous, can be used.

On the other hand, weighted log-rank tests are non-parametric methods which can also be used to compare two survival curves for time to event endpoints. Weighted log-rank test with the Fleming-Harrington class of weights can be used as the primary analysis in confirmatory studies of cancer vaccines focusing on a survival endpoint, with the purpose of avoiding a substantial loss of statistical power(Hasegawa, T., 2014). In contrary to the generalized pairwise comparisons, weighted log-rank tests can assess survival times for treatment groups for time event outcomes. In this case the hazard ratio is used as the measure of treatment effect. Hazards ratio interpretation can be meaningless or misleading when proportional hazards is assumed when hazards are indeed not proportional.

When using generalized pairwise comparisons, the net benefit is easier to interpret and has a direct link to individual patients. Survival differences by net benefit are more relevant to individual patients rather than using the hazard ratio which interprets the differences as an overall risk reduction. The net benefit is more advantageous in that a certain threshold is used to determine the benefit of the new treatment than the control. A treatment is considered beneficial if the survival differences exceeds a pre-specified clinically relevant threshold. The net benefit can be tested for significance using randomization tests.

Major advantage of generalized pairwise comparisons is that multiple outcomes can be analyzed simultaneously. Outcome measure are prioritized according to their clinical relevance. This applies in situations where more than one outcome measures are collected to answer the main objective. This provides much individual information in the assessment of benefit by a new treatment. There is no imputation required when using generalized pairwise comparisons. A pair of individual observations which are both censored are not classified as non-informative. An extension of the generalized pairwise comparisons proposed by Peron *e tal* 2016 uses the magnitude of the censored observations to calculate the score value which can be used in the calculation of the net benefit.

The weighted log-rank test has an advantage that it is flexible and can be adjusted for early, middle or late treatment effects. If the treatment group of interest is expected to have an effect earlier, a test which gives more weight to early failure times is pre-specified. The same is applicable for late and middle treatment effects. The problem is that it can be difficult to tell if the treatment is going to have an earlier or delayed treatment effect. The weighted log-rank test will lose power if the weights are not specified correctly. For example if it was pre-specified that with the new treatment under study there is going to be an early treatment effect and the new treatment effect started working late during follow up. The tests which detects early treatment effect will have little power to detect the delayed treatment effect thus the probability of false positive will be high. Thus the issue of pre-specifying the weights should be done with great caution because it might result in the loss of an effective treatment.

In the estimation of treatment effect, the net benefit depends on the threshold of clinical relevance. Under proportional hazards, the net benefit decrease with an increase in the threshold of clinical relevance. Large net benefits are observed when any treatments benefit is considered clinically relevant. In the presence of delayed treatment effect, the net benefit increase with an increase in the threshold of clinical relevance and decrease when long term thresholds are considered. The net benefit is is more pronounced when a proportion of patients is cured overally. The net benefit increase with an increase in the threshold of clinical relevance. The net benefit is even higher for long term thresholds. The same patterns are observed in the presence of censored observations. The net benefit is lower in the presence of censored observations as compared to situations where there is complete observations.

In this study the power of the weighted log-rank tests was compared to the power of generalized pairwise comparisons for different thresholds of clinical relevance. The power of the log-rank test which is a special case of the weighted log-rank tests depends only on the number of events since the hazards are assumed to be constant over time. The power of the weighted log-rank tests depends on the number of events and time of analysis but the power of generalized pairwise comparisons depends on the threshold of clinical relevance and time of analysis as well. The power of the log-rank test was higher than the power of the net benefit and other weighted log-rank tests under proportional hazards despite any threshold of clinical relevance. The net benefit lost power when the survival differences were getting larger(increasing $t$). However the power of the weighted log-rank tests are not affected by changes in the threshold of clinical relevance but by the time of analysis.

The log-rank test is sub-optimal when there is a delay in treatment effect for example treatment effect

of some cancer vaccines used for treating cancers. The log-rank tests lost power and even worst in the availability of censored observations. The net benefit gained power with increase in the survival differences and begin to lose power when very large survival differences are encountered. The Fleming and Harrington with $\rho = 0$ and $\gamma = 1$ has the highest power under delayed treatment scenario. It is as powerful as the net benefit for large thresholds for example $t = 42$ months. The net benefit loses power in the presence of censored observations especially when long thresholds are considered.

The power of net benefit is more defined when there is an overall cure in patients. It is more powerful than all weighted log-rank tests under cure rate scenario. It does not lose much power in the presence of censored observations. However the log-rank test and the Fleming and Harrington $G^{1,1}$ are least powerful when treatment effect is delayed and when there is overall cure in patients. The net benefit remains powerful under cure rate even if large survival differences are encountered. The net benefit is least powerful for very small survival differences for example $t = 2$ months. The net benefit depends on the threshold of clinical relevance and duration of follow-up time. The duration of trial follow-up time has a direct impact on censoring. However these issues are worth further research.

However these results are more meaningful if time of analysis is taken into account. It plays an important role and it is also related to the threshold of clinical relevance. Small thresholds imply short time follow up and very large thresholds imply a long time follow-up. If the treatment effect is delayed, a later analysis will have higher power to detect a given treatment effect than an earlier analysis with the same number of events. This means that the tests which gives much emphasis on the late failure times will have much power to detect treatment effect when analysis is done late and results can be otherwise if analysis is done early. For example in the simulation results under delayed treatment effect scenario, the net benefit and the Fleming and Harrington $G^{0,1}$ is more powerful if analysis is done late. If analysis is done early the net benefit might not have much power to detect treatment effect than other tests. If new treatment shows effects earlier, an analysis done earlier will also have greater power to detect treatment effect than late. This also means that tests which gives more emphasis to earlier failure times will have much power to detect treatment effect for example the Peto-Peto test and Gehan Generalized Wilcoxon test.

In the case of proportional hazards scenario, time of analysis will not have any impact since hazards are proportional over time thus the log-rank test will still remain the most powerful to detect treatment effect when hazards are proportional. Another interesting feature of the log-rank test is that the treatment effect is expressed as a hazard ratio, which is constant over time, that measures the relative benefit. However this property does not apply to weighted log-rank tests since they depend on the time of analysis. Under the cure rate scenario, some patients gets cured with time. Thus analyses done late will have much power to detect treatment effect. If an analysis is done late, the net benefit will still remain the most powerful under cure rate scenario. The issue of time of analysis is very important in weighing the power of the tests and might require further research.

# References


1. Buyse, M. (2010). Generalized pairwise comparisons for prioritized outcomes in the two sample problem. *Statist Med*, **29(30)**, 3245-3257.

2. Buyske, S., Fagerstrom, R. & Ying, Z. (2012). A Class of Weighted Log-Rank Tests for Survival Data When the Event is Rare. *Journal of the American Statistical Association*, **95(449)**, 249-258.

3. Fine, G D. (2007). Consequences of delayed treatment effects on analysis of time-to-event endpoints. *Drug Information Journal*, **41(4)**, 535-539.

4. Fleming, T R. & Harrington, D P. (1991). Counting Processes and Survival Analysis. *New York: John Wiley.*

5. Fleming, T R., Harrington, D P. & O'Sullivan, M. (1987). Supremum Versions of the Log-Rank and Generalized Wilcoxon Statistics. *J American Statistical Association*, **82(397)**, 312-320.

6. Gehan, A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika*, **52(1/2)**, 203-223.

7. Hasegawa, T. (2014). Sample size determination for the weighted log-rank test with the Fleming-Harrington class of weights in cancer vaccine studies. *Pharm Stat*, **13(2)**, 128-135.

8. Jacobs, I. (2016). Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): A randomized controlled trial. *The Lancet*, **387(10022)**, 945-956.

9. Karadeniz, P G. & Ercan, I. (2017). Examining tests for comparing survival curves with right censored data. *Statistics in Transition New Series*, **18(2)**, 311-328.

10. Lin, R S. & Leon, L F. (2017). Estimation of treatment effects in weighted log-rank tests. *Contemporary Clinical Trials Communications*, **8**, 147-155.

11. Mantel, N. & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *Journal of the National Cancer Institute*, **22(4)**, 719-748.

12. Ozenne, B. & Peron, J. (2016). Package buysetest. online. **https://mran.microsoft.com/web/packages/BuyseTest/BuyseTest.pdf**.

13. Peron, J., Buyse, M., Ozenne, B., Roche, L. & Roy, P. (2016). An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Statist Meth Med*, **27(4)**, 1230-1239.

14. Peron, J., Roy, P., Ozenne, B., Roche, L. & Buyse, M. (2016). The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials. *JAMA Oncol*, **2(7)**, 901-905.

15. Peto, R., & Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society*, **135(2)**, 185-207.

16. Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, **68**, 316-319.

17. Su, Z. & Zhu, M. (2017). Is it time for the weighted log-rank test to play a more important role in confirmatory trials? *Contemporary Clinical Trials Communications*, ISSN: 2451-8654.

18. Tarone, R E. & Ware, J. (1977). On Distribution-Free Tests for Equality of Survival Distributions. *Biometrika*, **64(1)**, 156-160.

19. Yang, S. & Prentice, R. (2010). Improved Log-rank-Type Tests for Survival Data Using Adaptive Weights. *Biometrics*, **66(1)**, 30-38.

# Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
**Statistical inference using generalized pairwise comparisons**

Richting: **Master of Statistics-Biostatistics**
Jaar: **2018**

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,

**Machingura, Abigirl**

Datum: **15/06/2018**