SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS
POSTGRADUATE PROGRAM

Methodological Issues with Proportional Hazard Models

By
Maria-Tereza Dellaporta

M.Sc. Thesis
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
February 2022

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Θέματα Μεθοδολογίας σε Μοντέλα Αναλογικών Κινδύνων

Μαρία-Τερέζα Δελλαπόρτα

Διατριβή
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Φεβρουάριος 2022

DEDICATION

To my beloved grandmother Chrysoula,
who was a constant inspiration

ACKNOWLEDGEMENTS

I would like to thank the following people for their unreserved support and continued help throughout the writing of my thesis.

First of all, I would like to say a special thanks to my thesis supervisor Prof. Dimitris Karlis for guiding me, advising me, and making time for my questions even when he had none. His clever remarks and brilliant suggestions have prompted me to improve the quality of my work, while his love for statistics is so inspiring that has considerably changed the way I approach every problem related to this field.

I would also like to express my sincerest gratitude to the Frontier Science Foundation team who has welcomed and supported me since day one. More specifically, I want to thank Prof. Urania Dafni for trusting me and providing me with a scholarship. I am extremely grateful for the opportunity to work for such a distinguished organization, and I sincerely hope this thesis will prove helpful in future problems related to the field of Biostatistics.

Additionally, many thanks go to my family and my friends who had and still have faith in me. My mother, Mary Mavroyanni, and my grandmother, Chrysoula Mavroyanni, have played a major role in helping me maintain my sanity throughout my academic years, and for that and many more, I am deeply grateful to them. Finally, I would like to express my love and appreciation to my closest friend, Alexia, and my partner Vasilis, for supporting me and easing my mind through difficult times.

ABSTRACT

Methodological Issues with Proportional Hazard Models

by Maria-Tereza Dellaporta

February 2022

A big part of survival data analysis is mainly based on two well-known methods: the log-rank test for the comparison of survival curves and the Cox proportional hazard model for the estimation of the effect corresponding to numerous variables of interest. Both methods are based on the assumption of proportional hazards. Due to the popularity, usefulness, and computational simplicity of these approaches, a potential violation of the proportional hazards assumption, which is an essential property for the validation of their findings, is oftentimes overlooked. In recent years, non-proportional data are frequently encountered, especially in the field of Biostatistics, where clinical trial data exhibit irregular patterns as a result of the administration of novel medicinal products and the implementation of innovative therapeutic procedures with unprecedented mechanisms of action.

To safeguard the validity and the generalizability of the results occurring from the analysis of such data, an in-depth literature review regarding various tests for the proportional hazards assumption and numerous testing procedures for the significance of treatment effect, in the two-sample case, is presented in this dissertation. Alternative modeling approaches and summary measures for the treatment effect are also discussed briefly, and an intuitive interpretation of the constant hazard ratio estimated via Cox's partial likelihood is given when the proportionality assumption is invalid. Two simulation studies, one for each group of tests, are conducted under proportionality and four non-proportional hazard patterns usually reported in contemporary publications.

Amongst the eighteen tests for proportionality examined, three of them display stable behavior under dissimilar types of departure from the null hypothesis: Grambsch & Therneau's suggestion (1994) using as functions of time either the ranks of the failure times or the Kaplan–Meier estimate of the pooled survivor function, and a modification of the goodness-of-fit test proposed by Lin (1991) using as weighted parameter estimators the ones introduced by Schemper et al. (2009). On the other hand, the comparison of twenty tests for treatment effect shows the superiority and flexibility of a newly developed method, which also provides piecewise hazard ratio estimates, called the Cauchy combination of change-point Cox regressions (Zhang et al., 2021). At the same time, various versatile weighted log-rank tests achieve good power under all hypothetical scenarios, except for the case of crossing hazards, where

the joint test by Royston & Parmar (2014) and a combination testing procedure by Breslow et al. (1984) noticeably surpass the other choices, in terms of performance.

In conclusion, the pattern of non-proportionality is definitive for the statistical analysis plan of time-to-event data. The optimal method, both for testing the assumption of proportional hazards and the significance of the treatment effect, is trial-specific. Nevertheless, when no a-priori knowledge exists about the anticipated type of non-proportionality, the aforesaid approaches seem to have good properties and are suggested for future analyses, until better methods arise.

ΠΕΡΙΛΗΨΗ

Θέματα Μεθοδολογίας σε Μοντέλα Αναλογικών Κινδύνων

Μαρία-Τερέζα Δελλαπόρτα

Φεβρουάριος 2022

Ένα μεγάλο μέρος της ανάλυσης δεδομένων επιβίωσης στηρίζεται σε δύο πολύ γνωστές μεθόδους: στον έλεγχο log-rank για τη σύγκριση δύο καμπυλών επιβίωσης, και στο μοντέλο αναλογικών κινδύνων του Cox για την εκτίμηση της επίδρασης διάφορων μεταβλητών ενδιαφέροντος. Λόγω της κοινής τους αποδοχής, της χρησιμότητας και της ευκολίας εφαρμογής τους, μία πιθανή παραβίαση της υπόθεσης των αναλογικών κινδύνων, η οποία αποτελεί βασική προϋπόθεση για την εγκυρότητα των αποτελεσμάτων τους, συχνά παραβλέπεται. Η συχνότητα εμφάνισης τέτοιων δεδομένων έχει αυξηθεί ραγδαία. Ιδιαίτερα, στον τομέα της Βιοστατιστικής, η χορήγηση νέων φαρμάκων και η εφαρμογή καινοτόμων θεραπειών οδήγησαν τα τελευταία χρόνια σε απρόβλεπτες δομές δεδομένων λόγω των πρωτοφανών μηχανισμών αλληλεπίδρασης τους με τον ανθρώπινο οργανισμό.

Για να εξασφαλιστεί η εγκυρότητα και η γενικευσιμότητα των ευρημάτων που προκύπτουν από την ανάλυση τέτοιων δεδομένων, μια λεπτομερής ανασκόπηση της υπάρχουσας βιβλιογραφίας, όσον αφορά ελέγχους για την υπόθεση των αναλογικών κινδύνων και τη στατιστική σημαντικότητα της επίδρασης μίας θεραπείας, παρουσιάζεται στην παρούσα διπλωματική εργασία. Επιπλέον, γίνεται μία σύντομη εισαγωγή σε εναλλακτικές προσεγγίσεις μοντελοποίησης και συνοπτικά μέτρα για την επίδραση της θεραπείας υπό μελέτη, και ταυτόχρονα δίνεται μία διαισθητική ερμηνεία στην εκτίμηση του λόγου κινδύνου που προκύπτει από το μοντέλο του Cox όταν δεν ισχύει η υπόθεση των αναλογικών κινδύνων. Δύο μελέτες προσομοίωσης, μία για κάθε ομάδα ελέγχων, διεξάγονται υπό την υπόθεση της αναλογικότητας αλλά και για τέσσερις περιπτώσεις μη αναλογικών κινδύνων που συχνά αναφέρονται στη σύγχρονη βιβλιογραφία.

Ανάμεσα στους δεκαοκτώ ελέγχους που έγιναν για την υπόθεση των αναλογικών κινδύνων, τρεις από αυτούς παρουσιάζουν σταθερή συμπεριφορά ανεξαρτήτως του βαθμού απομάκρυνσης από τη μηδενική υπόθεση ή το εναλλακτικό σενάριο: δύο από αυτούς ανήκουν στην οικογένεια ελέγχων των Grambsch & Therneau (1994) και χρησιμοποιούν ως συναρτήσεις του χρόνου είτε τον συνολικό εκτιμητή κατά Kaplan-Meier της συνάρτησης επιβίωσης ή την τάξη των χρόνων αποτυχίας, ενώ η τρίτη μέθοδος αποτελεί μία τροποποίηση του ελέγχου καλής προσαρμογής του Lin (1991), με σταθμισμένες εκτιμήσεις παραμέτρων αυτές που παρουσιάστηκαν στη σχετική δημοσίευση των Schemper, Wakounig και Heinze (2009). Από την άλλη μεριά, η σύγκριση είκοσι ελέγχων για τη στατιστική σημαντικότητα της επίδρασης μίας θεραπείας, δείχνει την

ανωτερότητα και την ευελιξία μιας νέας μεθόδου, που συγχρόνως παρέχει κατά τμήματα σταθερούς εκτιμητές για τον λόγο κινδύνου, και είναι γνωστή, εν συντομία, ως Cauchy CP (Zhang κ.ά., 2021). Συγχρόνως, μια ποικιλία ευέλικτων σταθμισμένων ελέγχων log-rank παρουσιάζουν καλά επίπεδα ισχύος για όλα τα προσομοιωμένα σενάρια, εκτός από εκείνο των διασταυρωμένων συναρτήσεων επιβίωσης, όπου την καλύτερη επίδοση, εμφανίζουν δυο μέθοδοι που συνδυάζουν τον έλεγχο log-rank με έναν έλεγχο για την υπόθεση της αναλογικότητας.

Συμπερασματικά, η φύση της μη αναλογικότητας είναι αυτή που καθορίζει το πλάνο της στατιστικής ανάλυσης των δεδομένων επιβίωσης. Η βέλτιστη μέθοδος, είτε για τον έλεγχο της υπόθεσης των αναλογικών κινδύνων, είτε για τη σημαντικότητα της επίδρασης της θεραπείας στην επιβίωση των ασθενών, εξαρτάται από την εκάστοτε κλινική δοκιμή. Ωστόσο, όταν δεν υπάρχει κάποια πληροφορία σχετικά με τη συμπεριφορά της συνάρτησης του λόγου κίνδυνου, οι παραπάνω μέθοδοι φαίνεται να έχουν καλές ιδιότητες και προτείνονται για μελλοντική χρήση, μέχρις ότου να αντικατασταθούν από νέες, καλύτερες προτάσεις.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Contents

List of Tables


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

List of Figures


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 1

Introduction

1.1 Motivation of the thesis

Throughout the years, the cornerstone of the statistical analysis of survival data is the assumption of proportional hazards. This becomes apparent when one realizes that the two most popular techniques in the discipline of Biostatistics are the log-rank test for hypothesis testing and the Cox Proportional Hazards (PH) Model for treatment effect estimation. Both of them gained momentum due to their simplicity and the interpretability of their results. They achieve maximum power under the proportionality assumption, but when that is not the case, biased results may occur, distorting the findings of a clinical trial and jeopardizing its success. The majority of the clinical trials are designed according to these techniques, with a target hazard ratio (HR) in mind. When the data are collected, firstly, the log-rank test is used for a preliminary analysis. For instance, it is used to test if there is a significant treatment effect between two patient groups: the ones taking the placebo and the ones receiving a new therapy. Since it is impossible to simultaneously adjust for many covariates using only the log-rank test, the Cox model is implemented to carry out a multivariate analysis. It was not until recent years that statisticians started noticing patterns of non-proportional hazards more and more frequently, leading them to the realization that they must change their – design and analysis – approach. But, why now? Why are there so many instances of non-proportional hazards? And how was this problem tackled in the past?

A substantial departure from the PH assumption has been a common observation in the development of oncology drugs in recent years, with the emergence of targeted therapies and cancer immunotherapies. The corresponding trials, where patterns of either a delayed improvement in the intervention group or reverse of treatment effect throughout follow-up often occur, made clear the urgent need for a

different methodology, one whose credibility is not entirely based on the proportionality assumption. Of course, the biological revolution does not uniquely account for the increasing amount of non-proportional hazards patterns. It is known that we have entered the big data era and as a consequence, larger trials are being carried out in the last decade or so. It is easier to detect a departure from proportionality as more data result in increased power of the corresponding tests.

Nevertheless, non-proportionality is not something new. Various methods have been formed and applied ever since the introduction of the aforementioned approaches. Weighted log-rank tests, stratification, and time-varying coefficients are only a few examples of such methods. However, each alternative has a downside, especially when it is inappropriately implemented. Biostatisticians must be aware of the dangers and the traps that each approach has in store. At the same time, they need to keep in mind the research question. For example, is a clinical study conducted with the aim of determining which treatment is better or to get more in-depth knowledge on how a new intervention works? Is interpretability of the results important and if so, which factors must be carefully considered? Is the objective of the study a clear-cut answer or just a prediction? All these questions and many more should be taken into account not only in the analysis but also during the design of a clinical trial.

Unfortunately, even though many papers have been written regarding the problem of non-proportionality, there is not a well-structured methodology. Some attempts have been made towards this cause: in recent papers, especially in the ones written after 2000, many researchers overexerted themselves running numerous simulations of possible non-proportional hazards patterns, with various sample sizes so as to compare old and new tests for proportionality and treatment effect. Notwithstanding the large number of interesting and useful conclusions drawn till now, the relevant information is in disarray. It is crystal-clear that there is not a panacea and each problem should be tackled according to its special characteristics, but a good statistician must be aware of the possible solutions, their advantages and disadvantages, along with their superiority compared to other methods. It is impossible to do so considering the number of suggested methodologies, especially when it is evident that some of them are in disagreement with others.

This thesis attempts to offer both the theoretical background needed and a variety of simulation results from several hypothetical non-proportionality patterns, frequently encountered in clinical trials. The objective here is to provide a reviewed collection of analysis methods, focusing on statistical tests, under different types

of departure from the proportionality assumption. The theory and mathematical justification of each test along with its performance under numerous scenarios will help the readers obtain a critical perspective and be able to settle on a plan of action when they face similar problems. Although the ideal ultimate goal would be to lay the foundation for a proposed "common approach", this thesis emphasizes more on equipping the interested parties with skills and knowledge that will ensure the quality and correctness of future research findings.

1.2 Brief structure of the thesis

A plethora of techniques related to the statistical analysis of time-to-event (TTE) data, where a violation of the PH assumption is speculated, is to be presented in the current document to achieve the abovementioned objectives.

The main body of the thesis begins in Chapter 2, which is introductory and specifies the notation and the fundamental terminology used in Survival Analysis. Terms frequently encountered in this field are being clarified and a handful of some rather enlightening examples are being presented. The key elements here are concepts such as censoring, survival function, Kaplan–Meier estimator, the log-rank test, as well as the Cox PH model and its famous partial likelihood.

After the short introduction to Survival Analysis, Chapter 3 starts with the presentation of the four most common non-proportional hazards patterns found in the literature: early/diminishing effect, late/delayed effect, crossing hazards, and long-term survivors. The purpose of this presentation is to offer the readers a less vague perception of what non-proportionality is. It also prepares them for the subsequent sections of this chapter, which consist of several testing procedures, both statistical and graphical, regarding the PH assumption. Graphical tests are presented briefly, while formal statistical tests are thoroughly explored and justified by – an outline of – their proof.

Chapter 4 includes a simulation study based on numerous scenarios of non-PH. The objective of this chapter is to compare a considerable amount of tests for proportionality under different types of departure from this hypothesis. Only the two-sample case (for instance, intervention versus placebo group) is examined as it is the basis of any further analysis and usually, it is the most important issue we need to deal with in practice.

Next, another essential group of techniques is being presented: a great variety of tests for treatment effect. Approaches related to the Restricted Mean Survival

Time (RMST) difference, weighted Cox regression, and variants or combinations of weighted log-rank tests, are only a few of the methods reported and clarified in Chapter 5. Moreover, an intriguing interpretation of the hazard ratio and other measures of treatment effect linked to the previously mentioned tests is given here.

Once again, Chapter 6 includes a simulation study similar to the one conducted in Chapter 4. Nonetheless, it refers to the tests of treatment effect mentioned in Chapter 5. Comparisons of power and type I error are performed under a plethora of non-PH scenarios.

Finally, Chapter 7 summarizes the most notable findings of Chapters 3 to 6. It also paves the way for discussion and direct proposals for further research, sharing both concerns and benefits of the current thesis.

Chapter 2

Survival Analysis 101

2.1 Fundamental definitions and notation

2.1.1 What is survival analysis?

Statistical analysis takes many forms depending on the nature of the problem of interest. In medicine, economics, engineering, and other disciplines the focus is usually on the expected duration of time until an event occurs (time-to-event data). Specific techniques of analysis have been formed to optimize the information being utilized in cases like these, especially due to their particularities. The domain of statistics involving these techniques, which examine and model the anticipated time until the occurrence of an event of interest, is called *survival analysis*.

Due to the widespread usage of methods employed by survival analysis across various scientific areas, there are several synonyms used. For instance, in engineering, survival analysis is usually referred to as "reliability theory", in economics as "duration modeling" and in sociology as "event history analysis". Even though the tools implemented are based on the same fundamental principles of survival analysis, the term differs depending on the type of event under the microscope. Some examples of events are:

- death (medicine),
- relapse/recurrence of a disease (medicine),
- infection (medicine),
- divorce (sociology), and
- malfunctioning of a device (engineering).

The definition of the event is crucial for the analysis and must be explicit. On many occasions, clarifications and specific instructions should be given. For instance, biological death is definite and thus, there is no need for further elucidation when it is defined as the event of interest. On the other hand, the malfunctioning of a device is not well-defined. Machines consist of many parts, but if one of them is missing or broken, it does not necessarily mean that the device will not function efficiently. Frequently, some parts are only decorative or they offer a rarely used extra capability. Is the malfunction referred to as a practical problem or a difference between the device and its original design? Undoubtedly, if the definition of the event of interest is ambiguous, the findings of the analysis will lack consistency.

Despite the concerns mentioned above, survival analysis is widely applicable because the definition of an event can be manifold and so, not only can we handle data from various fields of science, but we can also perform multiple analyses within the same data set using different definitions for the event under consideration. More specifically, survival analysis is normally used to:

- describe survival data (via Kaplan–Meier curves or measures such as median survival time, for instance),
- compare survival times among several groups of interest (typically via the log-rank test or its variants), and
- construct statistical models which help determine the magnitude of the effect of each variable, whether it is qualitative or quantitative, on survival. Models may be parametric or semi-parametric.

2.1.2 Survival time and censoring

Survival (or failure) time is defined as the duration of time from the beginning of the monitoring period until an event (failure) occurs. In the field of Biostatistics, survival time is defined as the duration of time from the beginning of follow-up until the outcome of interest occurs, which is usually death or relapse of disease. When the outcome is death, statisticians are interested in the overall survival time (OS) of the patients, who are called subjects of the study. In this case, time can be measured in years, months, weeks, or even days.

In a mathematical context, survival time is just a non-negative random variable denoted by T. T can either be discrete or continuous, but the notation and proofs provided in the following sections will only refer to the continuous case for reasons of

simplification. It should be noted that the proofs are similar when T is considered a discrete random variable.

When one has only partial information about the time to event, but the exact survival time is unknown, a key analytical problem occurs called *censoring*. There are three basic types of censoring:

1. *Right censoring*: a subject is right-censored when the outcome of interest happens at some time point after the end of its follow-up period. This kind of censoring is the most frequent, especially in clinical trials, where, for instance, some patients drop out before the study ends and the events of others occur after the follow-up's termination.

2. *Left censoring*: a subject is left-censored when the outcome of interest happens at some time point before the start of its follow-up period. A brief example is a study in which the interest is focused on the age at which children learn a specific task. When the study begins some of them may already know how to perform this task (left censoring), while others may have not yet learned it by the end of their follow-up (right censoring).

3. *Interval censoring*: a subject is interval-censored if it is known that the event occurs between two times, but the exact time of failure is unknown. Here, an example is the detection of breast cancer via mammography, in women over the age of 40. Doctors recommend an annual examination and so, if cancer is detected, that means cancer cells started developing at some time point between two consecutive screenings.

When the mechanism determining the censoring distribution is out of the control of the researcher, the censoring is called *random* (e.g. lost to follow-up patients). Otherwise, it is called *fixed* (e.g. when the event of interest is death and a patient dies after the study ends). In particular, the right-censored observations that occur from the termination of the study period are the result of *administrative censoring*.

Finally, censoring is also divided into two subcategories according to its dependence on survival time. When there is no association between them, censoring is *independent*. For instance, in a clinical trial with a primary outcome of interest the OS of subjects in two treatment groups, with a predefined study period of three years, patients who die after the end of the study are considered as censored observations. However, this censoring appears because the researchers chose to monitor the subjects for three years and it has nothing to do with their health status. On

the other hand, patients who dropped out because they got sicker display a censoring status which is undoubtedly connected with their condition, and therefore their survival time. In that case, censoring is called *informative* because it contains information about the parameters characterizing the distribution of T.

When one analyzes survival data there is only information about the time each subject has been monitored before the occurrence of an event and an indicator that specifies whether this time represents the entire survival time T or its censored counterpart. If C is the censoring time, then the aforementioned indicator is defined as

$$\delta = \begin{cases} 1, & \textit{if } T \leq C \\ 0, & \textit{otherwise.} \end{cases}$$

The observed time of follow-up is always equal to or less than the actual time to event. It is in fact equal to the minimum of T and C.

2.1.3 Probability distribution of a survival random variable

There are several equivalent ways to characterize the probability distribution of a survival random variable. This can be done by using:

- The *density function* $f(t)$ if T is a continuous variable, which is defined as

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

  and the *probability mass function* $p(t) = P(T = t)$ if T is a discrete one.

- The *cumulative distribution function* $F(t)$ which corresponds to the proportion of individuals whose event occurred as a function of t, and is defined as

$$F(t) = P(T < t).$$

  For the continuous case $F(t) = \int_0^t f(u)du$.

- The *survivor function* $S(t) = 1–F(t) = P(T \geq t)$ which gives the probability that a person survives longer than some specified time t. All survivor functions share the following theoretical properties:

  1. They are non-increasing. As time passes, the value of $S(t)$ remains the same or becomes smaller.
  2. Since time 0 is the starting point of the follow-up no one has gotten the event yet, and therefore $S(0) = 1$.

3. As t tends to infinity, the probability that a person survives longer than t tends to 0, or $\lim_{t\to\infty} S(t) = 0$.

- The *hazard function* $\lambda(t)$ *(or conditional failure rate)* which is mathematically defined as
$$\lambda(t) = \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}.$$
Intuitively, the hazard function gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. It is always a non-negative function with no upper bound.

While the relationship among density, cumulative distribution and survivor function is obvious, their connection with the hazard function is not. Using the definition of conditional probability, the connection becomes apparent, because
$$\begin{aligned}\lambda(t) &= \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} \\ &= \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t \cdot P(T \ge t)} \\ &= \frac{1}{P(T \ge t)} \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}\end{aligned}$$
and therefore,
$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{2.1}$$

- The *cumulative (or integrated) hazard function* $\Lambda(t)$ which is defined as
$$\Lambda(t) = \int_0^t \lambda(u)du.$$
The cumulative hazard at time t equals the area under the hazard curve up to time t. A cumulative hazard curve shows the (cumulative) probability that the event of interest has occurred up to any point in time.

By employing (2.1), it appears that
$$\begin{aligned}\lambda(t) &= -\frac{d}{dt} \ln S(t) \Rightarrow \\ \Lambda(t) &= -\ln S(t)\end{aligned}$$
and thus,
$$S(t) = e^{-\Lambda(t)}. \tag{2.2}$$

If one of the previous functions is known, so are all the others. These relationships led to the definition of several models which can be fitted to the data under examination. For instance, some models assume a particular form of the hazard function or the cumulative hazard function (e.g. flexible parametric models suggested by Royston & Parmar (2002)), while others suggest that survival time follows a certain distribution (e.g. Accelerated Failure Time models). There are also various non-parametric methods, such as the Kaplan–Meier estimator, which provide a specific formula for the calculation of the survival function. Whether someone delineates the hazard or the survivor function directly via a model, the rest of the functions described above can be estimated too.

2.1.4 The assumption of proportional hazards

Now that the hazard function is defined, a formal definition for the proportionality assumption must be given as well:

> *The assumption of proportional hazards holds when the hazard ratio comparing any two specifications of predictors is constant over time. Equivalently, this means that the hazard for one observation is proportional to the hazard for any other observation in the data, where the proportionality constant is independent of time.*

In a clinical trial context, for instance, there are usually two groups of patients: the control and the intervention group. If, after the analysis, the estimated hazard ratio of death for the control compared with the intervention group is equal to 2.8, that means that the hazard for a person in the control group is approximately three times the hazard for a person who received treatment. Of course, this assumption is not always met and in the last few years, it seems to be rather unrealistic. For example, it is sensible to think that if the intervention group underwent surgery for tumor removal, the risk of death will be higher at the beginning of their follow-up, but as time passes, the survivors are anticipated to show substantial improvement in comparison to cancer patients at the control group. As a result, the hazard ratio of two individuals belonging to different treatment groups is not constant over time. In fact, according to what was said above, the hazard ratio of death for the control compared with the intervention group should increase as the study progresses (since the hazard for the patients who underwent surgery diminishes over time and the hazard for the control group remains constant). Fortunately, cases like this led to

the invention of many statistical tests that properly assess the validity of the PH assumption. These tests will be extensively studied in Chapters 3 and 4.

2.1.5 Widely used distributions in survival analysis

When it to comes to modeling or simulating survival data, numerous probability distributions for the survival time, as well as the censoring time, can be considered as suitable candidates. Time is always positive and thus, the distribution under consideration must correspond to a non-negative random variable which is usually assumed to be continuous. Examples of such distributions are the following:

- exponential distribution,
- gamma distribution,
- Weibull distribution,
- log-normal distribution,
- log-logistic distribution,

and many more, including their mixtures.

Since a burning issue of this thesis is the assumption of proportional hazards, further insight into the exponential and the Weibull distribution will be offered. Provided certain conditions are met, these two distributions ensure that proportionality of hazards holds for two or more groups of observations sharing the same covariate values. Additionally, the piecewise exponential distribution will be studied as well, as it is considered to mimic observed trial results quite closely (Lin et al., 2020) and will be used for the simulation studies in Chapters 4 and 6.

Exponential distribution

If survival time $T \sim Exp(\lambda)$, then for any $t \geq 0$,

$$f(t) = \lambda e^{-\lambda t},$$

$$F(t) = 1 - e^{-\lambda t},$$

$$S(t) = 1 - F(t) = e^{-\lambda t},$$

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda,$$

$$\Lambda(t) = \int_0^t \lambda(u)du = \int_0^t \lambda du = \lambda t.$$

It should be noted that λ is a positive number, often called the *rate parameter*. This rate is essentially the hazard rate, the potential for failing at time t, given that the event has not occurred until then. Since the hazard rate equals a constant λ, the risk of failing does not change over time.

In this case, the PH assumption always holds. Indeed, for two groups of observations where $T_1 \sim Exp(\lambda_1)$ for the first group and $T_2 \sim Exp(\lambda_2)$ for the second, the hazard ratio for any two individuals is:

$$\frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\lambda_2}{\lambda_1}$$

which is independent of time.

Weibull distribution

If survival time $T \sim Weibull(\lambda, p)$, with λ and p being positive numbers, then according to the usual parameterization reported in medical statistics, for any $t \geq 0$,

$$f(t) = \lambda p t^{p-1} e^{-\lambda t^p},$$

$$F(t) = 1 - e^{-\lambda t^p},$$

$$S(t) = 1 - F(t) = e^{-\lambda t^p},$$

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\lambda p t^{p-1} e^{-\lambda t^p}}{e^{-\lambda t^p}} = \lambda p t^{p-1},$$

$$\Lambda(t) = \int_0^t \lambda(u)du = \int_0^t \lambda p u^{p-1} du = \lambda t^p.$$

The numbers λ and p are called the *scale* and the *shape parameter* of the distribution, respectively. Typically, properties and special characteristics of this distribution are displayed via the following categorization:

1. When $p = 1$, Weibull reduces to an exponential distribution. This means the hazard rate remains constant over time.

2. When $p > 1$, the hazard function is increasing over time (*increasing Weibull model*). An instance here could be a group of patients who do not respond to treatment and as their disease progresses, the instantaneous potential of dying becomes higher.

3. When $0 < p < 1$, the hazard function is decreasing over time (*decreasing Weibull model*). Decreasing hazard rates are often for patients who underwent surgery.

In this case, the PH assumption holds only when p is the same across all groups of interest. If $T_1 \sim Weibull(\lambda_1, p_1)$ for one group and $T_2 \sim Weibull(\lambda_2, p_2)$ for another, the hazard ratio for the second group compared with the first is:

$$\frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\lambda_2 p_2 t^{p_2-1}}{\lambda_1 p_1 t^{p_1-1}}$$

which is independent of time if and only if $p_1 = p_2$. Otherwise, the HR is either an increasing ($p_2 > p_1$) or a decreasing function ($p_2 < p_1$).

Piecewise exponential distribution

When examining the exponential distribution, it was stressed that the hazard rate does not change over time. Unfortunately, this is rarely the case. Frequently, the hazard rate differs from one time period to another, and in real-life applications it does not necessarily have a particular smooth curve. A problem might be more complex than one a simple distribution, such as Weibull, can describe. It is noteworthy, considering the three possible situations outlined above, that the Weibull distribution allows for either a constant or a monotonic hazard function. What will happen if the hazard rate is increasing at the beginning of a study and decreasing at the end or vice versa?

The piecewise exponential distribution provides a more flexible approach. It assumes that the hazard rate remains constant within some specified time intervals. But how does this offer greater flexibility? Well, let's think about it: if time intervals are narrow enough, the hazard rate at the beginning and the end of each will not show a substantial difference. Therefore, it is sensible to assume that it is constant over each small time period. It will be shown later that similar reasoning was used by Andersen (1982) to create a statistical test for the validity of the PH assumption.

A general form of the hazard function of a piecewise exponential distribution with k change points is given below:

$$\lambda(t) = \begin{cases} \lambda_0, & if\ 0 < t \leq \tau_1 \\ \lambda_1, & if\ \tau_1 < t \leq \tau_2 \\ \vdots & \\ \lambda_{k-1}, & if\ \tau_{k-1} < t \leq \tau_k \\ \lambda_k, & if\ t > \tau_k \end{cases}$$

where $\lambda_i > 0, \forall i \in \{0, 1, 2, \ldots, k\}$. When there is only one change point τ_1, and consequently, two rates, λ_0 before τ_1 and λ_1 after, it holds that

$$\lambda(t) = \begin{cases} \lambda_0, & \textit{if } 0 < t \leq \tau_1 \\ \lambda_1, & \textit{if } t > \tau_1. \end{cases}$$

Thus, if $0 < t \leq \tau_1$

$$\Lambda(t) = \int_0^t \lambda_0 du = \lambda_0 t,$$

whereas if $t > \tau_1$

$$\Lambda(t) = \int_0^{\tau_1} \lambda_0 du + \int_{\tau_1}^t \lambda_1 du = \lambda_0 \tau_1 + \lambda_1 (t - \tau_1)$$

and consequently,

$$\Lambda(t) = \begin{cases} \lambda_0 t, & \textit{if } 0 < t \leq \tau_1 \\ \lambda_0 \tau_1 + \lambda_1 (t - \tau_1), & \textit{if } t > \tau_1. \end{cases}$$

From the formula **??**,

$$S(t) = e^{-\Lambda(t)} = \begin{cases} e^{-\lambda_0 t}, & \textit{if } 0 < t \leq \tau_1 \\ e^{-\lambda_0 \tau_1 + \lambda_1 (t - \tau_1)}, & \textit{if } t > \tau_1. \end{cases}$$

Therefore

$$F(t) = 1 - S(t) = \begin{cases} 1 - e^{-\lambda_0 t}, & \textit{if } 0 < t \leq \tau_1 \\ 1 - e^{-\lambda_0 \tau_1 + \lambda_1 (t - \tau_1)}, & \textit{if } t > \tau_1. \end{cases}$$

and

$$f(t) = \frac{d}{dt} F(t) = \begin{cases} \lambda_0 e^{-\lambda_0 t}, & \textit{if } 0 < t < \tau_1 \\ \lambda_1 e^{-\{\lambda_0 \tau_1 + \lambda_1 (t - \tau_1)\}}, & \textit{if } t > \tau_1. \end{cases}$$

For two groups of observations with hazard functions corresponding to a piecewise exponential distribution with k change points, the proportionality assumption is valid if and only if the change points are the same for the two groups and the HRs are equal across all time intervals. Thus, here the PH assumption holds under some strict conditions which are hardly met in practice. Notwithstanding this realization, piecewise exponential models are extremely useful for simulation studies and further investigation of statistical tools currently available.

2.2 Kaplan-Meier estimator

Probably the most popular approach for estimating a survivor function without resorting to parametric methods is the Kaplan–Meier (KM) estimator, frequently mentioned as the *product limit estimator*. Recalling that the survivor function $S(t)$ is just the probability an individual survives longer than or at least for a specified time t, results in a rather obvious empirical survival function, under the condition there is no censoring:

$$S(t) = \frac{\#\text{ of individuals with } T \geq t}{\text{total sample size}}.$$

However, survival data are data often occurring from a long period of subjects' monitoring. It is not unusual for information regarding the event of interest to get lost due to various reasons, examples for which have been given throughout the preceding sections. Thankfully, Kaplan & Meier (1958) proposed a way to non-parametrically estimate $S(t)$, even in the presence of censoring. The method is based on the fundamental concept of conditional probability and a well-known relevant law, the multiplication law of probability. According to this, for m events $A_1, A_2, \ldots, A_m$, it holds that

$$P(A_1 \cap A_2 \cap \ldots \cap A_m) = P(A_1)P(A_2|A_1)\ldots P(A_m|A_1, A_2, \ldots, A_{m-1}). \tag{2.3}$$

When a survival analysis is being conducted, the available data include information regarding the time of event or censoring and also, an event or censoring indicator to distinguish the observations providing a complete profile from those who offer a partial one. Let's consider the following notation: $t_1, t_2, \ldots, t_m$ are the exact ordered times at which one or more events occurred, r_j is the number of individuals at risk at time t_j, meaning that they at least survived until then and it is possible to "fail" at t_j or in the future, while d_j is the number of failures at t_j. Then for $t \in [0, t_1]$ it is reasonable to estimate $S(t)$ as the proportion of individuals who survived at least until t_1, and thus,

$$\hat{S}(t) = P(T \geq t_1) = 1, \text{for } t \in [0, t_1].$$

In general, for $t \in (t_k, t_{k+1}]$,

$$\begin{aligned}
\hat{S}(t) &= P(T \geq t_{k+1}) \\
&= P(T \geq t_1, \ldots, T \geq t_k, T \geq t_{k+1}) \\
&\overset{(2.3)}{=} P(T \geq t_1) \prod_{j=1}^{k} P(T \geq t_{j+1} | T \geq t_j) \\
&= \prod_{j=1}^{k} \{1 - P(T = t_j | T \geq t_j)\}
\end{aligned}$$

since $P(T \geq t_1) = 1$. It is rational to estimate $P(T = t_j | T \geq t_j)$ as $\frac{d_j}{r_j}, \forall j \in \{1, 2, \ldots, m\}$. Consequently,

$$\hat{S}(t) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{r_j}\right), \text{for } t \in (t_k, t_{k+1}], k \in \{1, 2, \ldots, m\} \tag{2.4}$$

The estimate is a left continuous step function whose value changes only shortly after an event occurs. Nevertheless, it should be noted that slight deviations in the notation and the definition of the survivor function can lead to a similar, but right continuous Kaplan–Meier estimator, meaning that the change in its value happens exactly when an event takes place. Finally, it should also be noted that the KM method assumes censoring is independent of survival time, or put in simple words, the reason an observation is censored is unrelated to the cause of failure.

2.3 Log-rank test

The comparison of two or more survival curves has always been an important problem in survival analysis. Numerous parametric and non-parametric methods have been developed, some for censored and others for uncensored observations. Focusing on non-parametric approaches, the log-rank test is widely accepted as the most famous amongst the available options, especially for censored data.

For only two populations, the null hypothesis is

$$H_0 : S_1(t) = S_2(t),$$

meaning that the burning issue is whether the distributions of survival times in the two groups are identical or not. The anticipated alternative would be:

$$H_A : S_1(t) \neq S_2(t).$$

However, the log-rank test achieves maximum power under the alternative of proportional hazards, i.e., $\lambda_1(t) = \theta\lambda_2(t)$ for some positive constant $\theta \neq 1$. Using (2.2), it is easy to conclude that the respective alternative hypothesis in terms of the survivor function is:

$$H_A : S_1(t) = [S_2(t)]^\theta.$$

Now that the statistical context is well-defined, it is time to present how the log-rank test works. The idea behind this test is based on the construction and combination of a sequence of 2×2 contingency tables displaying group versus survival status for each time t at which a failure occurs. The equivalent null hypothesis to the one mentioned before is that the survival profile is independent of the group. Once the entire sequence of 2×2 tables has been generated, the information contained in the tables is accumulated using one single statistic. This statistic compares the observed number of failures at each time to the expected number of failures given that the distributions of survival times for the two groups are identical. If the null hypothesis is true, the test statistic has an approximate chi-square distribution.

More specifically, at time t_j a contingency table similar to Table 2.1 can be constructed. If d_j, r_j and r_{1j} are regarded as fixed values, and the null hypothesis is true, d_{1j} can be considered as a random variable that follows the hypergeometric distribution, since when one subject in the first group fails it is impossible to fail again in the future[1] (sampling without replacement). The probability mass function for d_{1j} in this case is

$$p(d_{1j}|d_j, r_j, r_{1j}) = \frac{\binom{r_{1j}}{d_{1j}}\binom{r_j - r_{1j}}{d_j - d_{1j}}}{\binom{r_j}{d_j}}$$

with d_{1j}'s possible values ranging from 0 to $\min(d_j, r_{1j})$.

Group	No. of events	No. of survivors beyond t_j	Total
I	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
II	d_{2j}	$r_{2j} - d_{2j}$	r_{2j}
Total	d_j	$r_j - d_j$	r_j

Table 2.1: Log-rank test's contingency table at time point t_j.

It is easy to find that the mean of d_{1j} and thus, the expected number of failures at t_j is given by the formula

$$e_{1j} = d_j \frac{r_{1j}}{r_j}.$$

[1]We do not consider recurrent event survival analysis.

Similarly, the variance of d_{1j} is

$$v_{1j} = \frac{d_j r_{1j} r_{2j}(r_j - d_j)}{r_j^2(r_j - 1)}.$$

All m contingency tables corresponding to failure times $t_1, t_2, \ldots, t_m$ can be combined via the following statistic which depends on the difference between the observed and expected number of events:

$$X_{LR}^2 = \frac{U_1^2}{V_1} \tag{2.5}$$

where $U_1 = \sum_{j=1}^{m}(d_{1j} - e_{1j})$ and $V_1 = Var(U_1) = \sum_{j=1}^{m} v_{1j}$. Under H_0, the test statistic asymptotically follows a χ^2 distribution with 1 degree of freedom (d.f.).[2]

Despite the fact that the log-rank test is essentially testing the equivalence of two survival functions versus the alternative of proportional hazards, this does not mean its results are invalid when another alternative relationship holds. Unfortunately, however, there are some special cases (e.g. when the survivor functions cross each other) where the test lacks a great amount of power. In Chapters 5 and 6 more insight and ways to tackle this problem will be given both in a theoretical and a practical context.

2.4 Cox PH model

Fifty years ago, Sir David Cox (1972) made a groundbreaking proposal, that of a new statistical model, specially created for survival data. The proportional hazards regression model of Cox has since become the most known semi-parametric model for the analysis of failure time regression data. Today, the Cox model is used in countless applications, not only in survival analysis but also in related fields, such as reliability analysis, epidemiology, and biomedical studies. Apart from Statistics and Biostatistics, many other disciplines have also benefited, including Biology, Actuarial Science, and Finance.

But why is Cox's model so popular? What is the characteristic or the property that makes it remarkable? The following sections present the basic qualities of the Cox PH model, mainly from a mathematical point of view. Hopefully, by the end of this chapter, the reasons for the Cox model's popularity will become apparent.

[2]Interestingly, an approximation to the log-rank statistic can be calculated using observed and expected values for each time t_j without having to compute the variance formula. The approximate formula is of the classic chi-square form that sums the square of the observed minus expected value divided by the expected value over all failure times. Nevertheless, this approximation is not frequently used but employs the same rationale as the log-rank test (Kleinbaum et al., 2012).

2.4.1 Formula and partial likelihood of the model

Starting from a mathematical perspective, the formula of the Cox PH model says that the hazard for the *i-th* individual at time t is the product of two quantities. The first of these, $\lambda_0(t)$, is called the *baseline hazard* and it is an unspecified function. No parametric assumption is made for $\lambda_0(t)$ and that is why the model is semi-parametric, in comparison to parametric models with similar forms, such as the Weibull or the exponential where the baseline hazard is specified (according to Section 2.1.5). The second quantity is an exponential expression, independent of time t.

Let n be the number of individuals in the analysis with censoring or failure times $t_1, t_2, \ldots, t_n$, respectively. If p characteristics of the population are being under consideration, then denote by x_i the $p \times 1$ vector of predictor variables for subject $i, i = 1, 2, \ldots, n$, and β a $p \times 1$ vector of unknown regression parameters. Then, the hazard function for the failure time of the *i-th* individual is given by the following formula:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta' x_i). \tag{2.6}$$

In order to fit a Cox model to a data set, one must estimate the parameter vector β. The elements of the aforementioned vector can be estimated after maximizing the partial likelihood (PL) function of the Cox model. Interestingly, the PL was initially referred to as conditional likelihood by Cox (1972), but one year later, Kalbfleisch and Prentice's comments on his paper, led to the realization that this function was in fact a partial likelihood (Kalbfleisch & Prentice, 1973; Cox, 1975). Without further delay, let R_i be the set of individuals who have not failed or been censored by t_i (risk set at time t_i), and δ_i be the event indicator for subject i, meaning that $\delta_i = 1$ if subject i failed at t_i, and 0 otherwise. Under the condition there are no ties, i.e., at most one event occurred at each time t_i, the PL of the previous model is

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta' x_i)}{\sum_{\ell \in R_i} \exp(\beta' x_\ell)} \right]^{\delta_i} \tag{2.7}$$

Of note, censoring times are effectively excluded from the likelihood because for these observations the exponent δ_i equals 0. Also notable is that the ratio

$$\frac{\exp(\beta' x_i)}{\sum_{\ell \in R_i} \exp(\beta' x_\ell)}$$

has an intuitive interpretation: according to the proportional hazards model, the hazard for subject i, for whom the event was actually observed to occur at time

t_i, is proportional to $\exp(\beta' x_i)$. Consequently, this ratio expresses the hazard for subject i in relation to the cumulative hazard for the subjects at risk at time t_i. This explains why someone would choose to estimate β via finding the maximum point of $L(\beta)$. Each time t_i, for which $\delta_i = 1$, corresponds to the time of failure for a subject with label i. The probability of failure for this subject at that exact time should be higher than the respective probability for any other subject included in R_i. Therefore, the ratio

$$\frac{\exp(\beta' x_i)}{\sum_{\ell \in R_i} \exp(\beta' x_\ell)}$$

should be as high as possible. Since $L(\beta)$ is essentially the product of all these ratios, maximizing $L(\beta)$ will, in a sense, maximize each factor, while adjusting for the others, simultaneously.

Unfortunately, this maximization is not feasible by hand and there is not a closed-form solution. A usual method to deal with this issue is the root-finding algorithm Newton–Raphson. This algorithm has been developed to solve difficult equations, i.e., equations for which there is not a specific methodology, to begin with. However, here, no equation has been written. So, first of all, the partial log-likelihood must be defined:

$$\ell(\beta) = \ln L(\beta) = \sum_{i=1}^{n} \delta_i \{\beta' x_i - \ln \sum_{\ell \in R_i} \exp(\beta' x_\ell)\}. \tag{2.8}$$

After differentiating the above function with respect to β, it occurs that

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{\ell \in R_i} x_\ell \exp(\beta' x_\ell)}{\sum_{\ell \in R_i} \exp(\beta' x_\ell)} \right]. \tag{2.9}$$

Frequently, the derivative of $\ell(\beta)$ is denoted by $U(\beta)$, and the equation

$$U(\beta) = 0 \tag{2.10}$$

is called the *partial likelihood score equation*. Intuitively, $U(\beta)$ expresses the sum of the differences between observed and expected covariate values over the subjects who failed, since the term $\sum x_\ell \exp(\beta' x_\ell) / \sum \exp(\beta' x_\ell)$ is a weighted average of x_i over all individuals at risk at time t_i. The *maximum partial likelihood estimators* (MPLE) can be found by solving $U(\beta) = 0$, employing a computer program and of course, a root-finding algorithm. These estimators share the general properties of the maximum likelihood estimates.

2.4.2 Approximation methods for tied survival times

Luckily, a similar procedure can be followed when there are ties, i.e., there is at least one time point at which two or more events take place. Nevertheless, the

computation of the MPLE, in this case, is time-consuming. Several proposals have been made to circumvent this obstacle, with the most popular being Efron's and Breslow's adjustments for ties.

Breslow's modified partial likelihood

Breslow (1974) proposed an approximation of the exact PL when ties are present. Let $t_1 < t_2 < \ldots < t_m$ be the ordered failure times, d_j be the number of events at t_j, R_j the risk set at t_j, and finally, S_j the sum of the covariate values over all subjects who failed at t_j. Then, Breslow's modified PL is given by the formula

$$L(\beta) = \prod_{j=1}^{m} \frac{\exp(\beta' S_j)}{\left[\sum_{\ell \in R_j} \exp(\beta' x_\ell)\right]^{d_j}}. \tag{2.11}$$

This formula is just an approximation of Cox's discrete method for ties.

Efron's modified partial likelihood

Efron (1977) suggested a different method for the approximation of the PL. Using the same notation as before, and denoting by D_j the set of individuals who fail at t_j, the formula

$$L(\beta) = \prod_{j=1}^{m} \frac{\exp(\beta' S_j)}{\prod_{r=0}^{d_j-1} \left[\sum_{\ell \in R_j} \exp(\beta' x_\ell) - \frac{r}{d_j} \sum_{k \in D_j} \exp(\beta' x_k)\right]} \tag{2.12}$$

is quite close to the real PL.

When there are no ties, both methods give the same results as the initial PL presented in this section. However, under the presence of ties, other factors must also be considered regarding which approach should be implemented. Again, when there are few ties the results do not differ substantially. When their number is large, Breslow's approximation performs poorly, as it underestimates the regression parameters (the elements of β are biased towards 0), while Efron's method performs far better, even though estimators are biased too. Of course, other options are available, such as the discrete method by Cox (1972) or the exact method by Kalbfleisch & Prentice (1973). Exact methods yield more accurate results, but they are computationally demanding and time-consuming. Thus, in practice, either Breslow's or Efron's approximation is used. Although the Breslow approximation is the default in many standard software packages, the Efron method for handling ties is to be preferred, particularly when the sample size is small either from the outset or due to heavy censoring (Hertz-Picciotto & Rockhill, 1997).

2.4.3 PH assumption in Cox regression

Proportional hazards assumption is inseparably connected with the Cox model, at the extend that the latter is oftentimes referred to as Cox PH model or proportional hazards model. Indeed, along with its particular form, it's the only assumption a statistician should test to evaluate the validity and goodness of fit of the model. The definition of the Cox model makes evident that, for any two individuals i_1 and i_2 their hazard ratio

$$\frac{\lambda_{i_2}(t)}{\lambda_{i_1}(t)} = \frac{\lambda_0(t)\exp(\beta' x_{i_2})}{\lambda_0(t)\exp(\beta' x_{i_1})} = \frac{\exp(\beta' x_{i_2})}{\exp(\beta' x_{i_1})}$$

is independent of time. When the PH assumption is invalid, the estimates of the Cox model are biased and unstable. If a statistical analysis depends entirely on a Cox model, but hazards are not proportional, the findings are incorrect and misleading. Especially, in the field of Biostatistics where human lives are at stake, these mistakes should be avoided at all costs. In the next chapter, a variety of tests for proportionality will be presented with this issue in mind.

2.4.4 Association between the log-rank test and the Cox PH model

In the introduction of this thesis, it was mentioned that the log-rank test and Cox model are connected. Both statistical methods achieve their greatest performance under the PH assumption. However, this is not the only link between them: it can be proved that the log-rank statistic arises as a score test from the partial likelihood function.

The log-rank test is implemented for the comparison of groups into the data set. In the two-sample case, i.e., when there are only two groups the log-rank statistic is given by the formula described in section 2.3. Returning to the Cox model, let x be an indicator variable which is equal to 1 for the individuals belonging to the first group and 0 for the individuals belonging to the second group. Using only this covariate and the same notation as before, assuming that there are no ties for simplicity, the partial likelihood is

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta x_i)}{\sum_{\ell \in R_i} \exp(\beta x_\ell)} \right]^{\delta_i}.$$

The corresponding partial log-likelihood and its first and second derivatives are given below:

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i \{\beta x_i - \ln \sum_{\ell \in R_i} \exp(\beta x_\ell)\},$$

$$\ell'(\beta) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{\ell \in R_i} x_\ell \exp(\beta x_\ell)}{\sum_{\ell \in R_i} \exp(\beta x_\ell)} \right] \text{ and}$$

$$\ell''(\beta) = -\sum_{i=1}^{n} \delta_i \frac{\sum_{\ell \in R_i} {x_\ell}^2 \exp(\beta x_\ell) \cdot \sum_{\ell \in R_i} \exp(\beta x_\ell) - \left[\sum_{\ell \in R_i} x_\ell \exp(\beta x_\ell)\right]^2}{\left[\sum_{\ell \in R_i} \exp(\beta x_\ell)\right]^2}.$$

At this point, it is important to recall the null hypothesis of the log-rank test: the distributions of survival times in the two groups are identical. If this is the case, then the hazards for individuals of different groups will also be identical. This is equivalent to testing the hypothesis $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$, because under H_0 the HR of the two groups is equal to 1. The score test here is conducted using the statistic

$$\frac{U(\beta_0)^2}{I(\beta_0)}$$

where $U(\beta_0)$ is the partial likelihood score and $I(\beta_0)$ is the Fisher information, both evaluated at $\beta_0 = 0$. It is known that

$$\begin{aligned} U(\beta) = \ell'(\beta) \Rightarrow U(0) &= \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{\ell \in R_i} x_\ell \exp(0 \cdot x_\ell)}{\sum_{\ell \in R_i} \exp(0 \cdot x_\ell)} \right] \\ &= \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{\ell \in R_i} x_\ell}{|R_i|} \right] \\ &= \sum_{i=1}^{n} \delta_i \left[x_i - \frac{|R_{1i}|}{|R_i|} \right] \\ &= \sum_{i:\delta_i=1} \left[x_i - \frac{|R_{1i}|}{|R_i|} \right] \end{aligned}$$

where $|R_i|$ is the number of subjects at risk at t_i, and $|R_{1i}|$ is the number of subjects belonging to the first group at risk at t_i. It is obvious that $U(0) = U_1$ for the special case of no ties, where U_1 is the sum of observed minus expected number of events over all failure times.

Fisher information can be estimated via the negative of the second derivative of the partial likelihood (*observed Fisher information*), and thus

$$\begin{aligned} \hat{I}(\beta) &= -\ell''(\beta) \Rightarrow \\ J(0) = \hat{I}(0) &= \sum_{i=1}^{n} \delta_i \frac{\sum_{\ell \in R_i} {x_\ell}^2 \exp(0 \cdot x_\ell) \cdot \sum_{\ell \in R_i} \exp(0 \cdot x_\ell) - \left[\sum_{\ell \in R_i} x_\ell \exp(0 \cdot x_\ell)\right]^2}{\left[\sum_{\ell \in R_i} \exp(0 \cdot x_\ell)\right]^2} \\ &= \sum_{i:\delta_i=1} \frac{\sum_{\ell \in R_i} {x_\ell}^2 \cdot |R_i| - \left[\sum_{\ell \in R_i} x_\ell\right]^2}{|R_i|^2} \end{aligned}$$

but x's possible values are 1 and 0, so ${x_\ell}^2 = x_\ell \; \forall \ell$, and therefore,

$$\begin{aligned} J(0) &= \sum_{i:\delta_i=1} \frac{\sum_{\ell \in R_i} x_\ell \cdot |R_i| - \left[\sum_{\ell \in R_i} x_\ell\right]^2}{|R_i|^2} \\ &= \sum_{i:\delta_i=1} \frac{|R_{1i}| \cdot |R_i| - |R_{1i}|^2}{|R_i|^2} \\ &= \sum_{i:\delta_i=1} \frac{|R_{1i}| \cdot (|R_i| - |R_{1i}|)}{|R_i|^2} \end{aligned}$$

which is equivalent to V_1 when there are no ties.

Consequently, the score test statistic takes the form of the log-rank statistic, and like the latter, it follows a χ^2 distribution with 1 d.f. under H_0. This proves that the Cox model is just a generalization of the log-rank test for the multivariate case (Cox, 1972; Harrington, 2014).

2.4.5 Estimation of the baseline hazard

A frequent downside of the Cox model is that it only determines HRs, i.e., it gives answers regarding the relative risk between individuals with dissimilar characteristics, but not about their absolute hazards. This stems from the fact that the baseline hazard is unspecified. Fortunately, having estimated the parameters of a Cox model, it is possible to recover a non-parametric estimate of the baseline hazard function.

Of course, it is not necessary to fit a Cox model to gain a non-parametric estimate of the hazard or the cumulative hazard function. For instance, there is the well-known Nelson–Aalen estimator which is defined as

$$\hat{\Lambda}(t) = \sum_{j:t_j \leq t} \frac{d_j}{r_j} \tag{2.13}$$

where $t_1, t_2, \ldots, t_m$ are the failure times only, while d_j and r_j define the number of failures and subjects at risk at t_j, respectively. Another way to estimate the cumulative hazard is by obtaining the KM estimate of the corresponding survivor function and using (2.2) to estimate $\Lambda(t)$.

Nevertheless, none of the aforementioned approaches utilizes the findings of a fitted Cox model. Luckily, Breslow (1972) suggested estimating β and the baseline cumulative hazard $\Lambda_0(t)$ in the maximum likelihood framework. By treating $\lambda_0(t)$ as piecewise constant between uncensored failure times, one can show that the joint likelihood for β and Λ_0 is maximized simultaneously at $\hat{\beta}$, the maximum partial

likelihood estimator, and

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{\delta_i \cdot I(t_i \leq t)}{\sum_{\ell \in R_i} \exp(\hat{\beta}' x_\ell)}. \tag{2.14}$$

This is the well-known Breslow estimator. Like any other MLE, it is asymptotically normal and consistent. Its worth is apparent when one considers that all major statistical software packages, such as `SAS` and `R`, implement this formula for the estimation of the baseline cumulative hazard. At the same time, its strong presence in numerous scientific papers reflects the importance of using a non-parametric, Cox model-based estimator of the hazard function (see Chapter 3 for more).


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 3

Tests for proportional hazards

The typical analysis procedure for survival data starts with non-parametric methods, such as Kaplan-Meier estimates for the groups of interest, followed by the log-rank test (or some variant) for the comparison of the survival curves, and finally, the Cox model is fitted to provide an estimate of the magnitude of their difference. Under the assumption of proportional hazards, the results from the log-rank test and the Cox model are valid and informative of the nature of the data. However, nowadays non-proportionality of hazards is more and more frequently encountered, although it is ignored in many cases. For instance, most randomized clinical trials with a time-to-event outcome are designed assuming proportional hazards of the treatment effect, even though recent breakthroughs in the field of medicine showed that, depending on the mechanism of action, a therapy can display great effect in a non-conventional way, rather than in a consistent one. As a consequence, various patterns of non-PH are observed.

Patterns of non-PH may be obvious in the KM plots in the initial analysis. Nevertheless, when the number of groups under consideration is large or the variable of interest is continuous, these patterns are easily missed. For that reason, it is of great importance to check the PH assumption via strict statistical criteria. Numerous graphical and statistical tests have been developed throughout the years, mainly based on the Cox model. Usually, the graphical tests complement the statistical ones. The interpretation of graphs is subjective and thus, cannot be used alone to define whether the PH assumption is met or not. Instead, it is preferable to use statistical tests to offer a clear-cut answer regarding the validity of the proportionality of hazards, and then give a more intuitive interpretation of the result employing a relative plot, if possible. Towards this cause, this chapter begins with a small presentation of the most famous patterns of non-PH and continues with a wide variety of tests for the proportionality assumption, both statistical and graphical.

3.1 Frequent patterns of non-PH

Due to recent advancements in medical research, and specifically in oncology therapy, the proportionality assumption seems to be oftentimes violated. Several types of non-proportionality patterns usually occur either as a consequence of different treatment effects in subgroups or due to the treatment itself. Also, in recent years, the accumulation of data is larger, resulting in faster detection of existing non-PH patterns. While it is hard to specify every possible non-PH scenario, four types are repeatedly mentioned in the literature:

1. *Early/Diminishing effect*: With an early effect, the HR is significantly different from 1 in the early follow-up and approaches it as time passes. An early effect may, for example, be provoked by 'wearing off' of the effectiveness of a therapy that is administered for a limited period and then stopped.

2. *Late/Delayed effect*: This is the exact opposite of an early effect. At the beginning, the HR is close to 1 and as time passes their difference becomes larger and larger. Late effect can be observed when a treatment does not immediately improve the health status of the corresponding group but proves beneficial after some period of time. This is usually the case for immuno-oncology drugs, possibly due to their mechanism of action or due to the design of the trial (Ananthakrishnan et al., 2021). A delayed effect may also occur in screening[1] or prevention[2] trials, in which the treatment effect is expected to take time to manifest.

3. *Crossing hazards*: Sometimes, a short-term or a delayed benefit can also cause the hazard functions to cross each other. Another reason for crossing hazards stems from the fact that the treatment may be harmful in a subgroup but helpful in its complement. This phenomenon demands special consideration as the comparison of the treatment arms is not straightforward and a clear-cut answer for the superiority or inferiority of a remedy is not easily provided.

4. *Long-term survivors*: Finally, a fourth non-proportional hazards pattern observed in recent years is the one produced by long-term survivors. It is known

[1]Screening trials evaluate new tests for detecting cancer and other health conditions in people before symptoms are present. The goal is to determine whether the screening test saves lives and at what cost.

[2]Prevention trials involve tests to find ways to prevent particular medical conditions or if people have them already, to prevent them from reoccurring. The emphasis of these studies might be on medicines, vitamins, minerals, or lifestyle changes.

that therapies for certain cancer types are believed to induce a subset of long-term survivors, and in some diseases, normally a proportion of patients are expected to be cured (or non-susceptible), that is to remain alive or disease-free even after long follow-ups (Chen, 2013).

To have a better insight, Figure 3.1 displays the survivor functions for two groups in each scenario. It is vital to remember that these are only indicative and more complicated scenarios are encountered in real clinical data. Furthermore, we should keep in mind that when the Cox PH model is used to provide a hazard ratio estimation on these occasions, the resulting summary statistic may be under-or overestimated, while the traditional log-rank test lacks power. That is why it is crucial to test the PH assumption before reporting findings based on the Cox regression and the log-rank test.

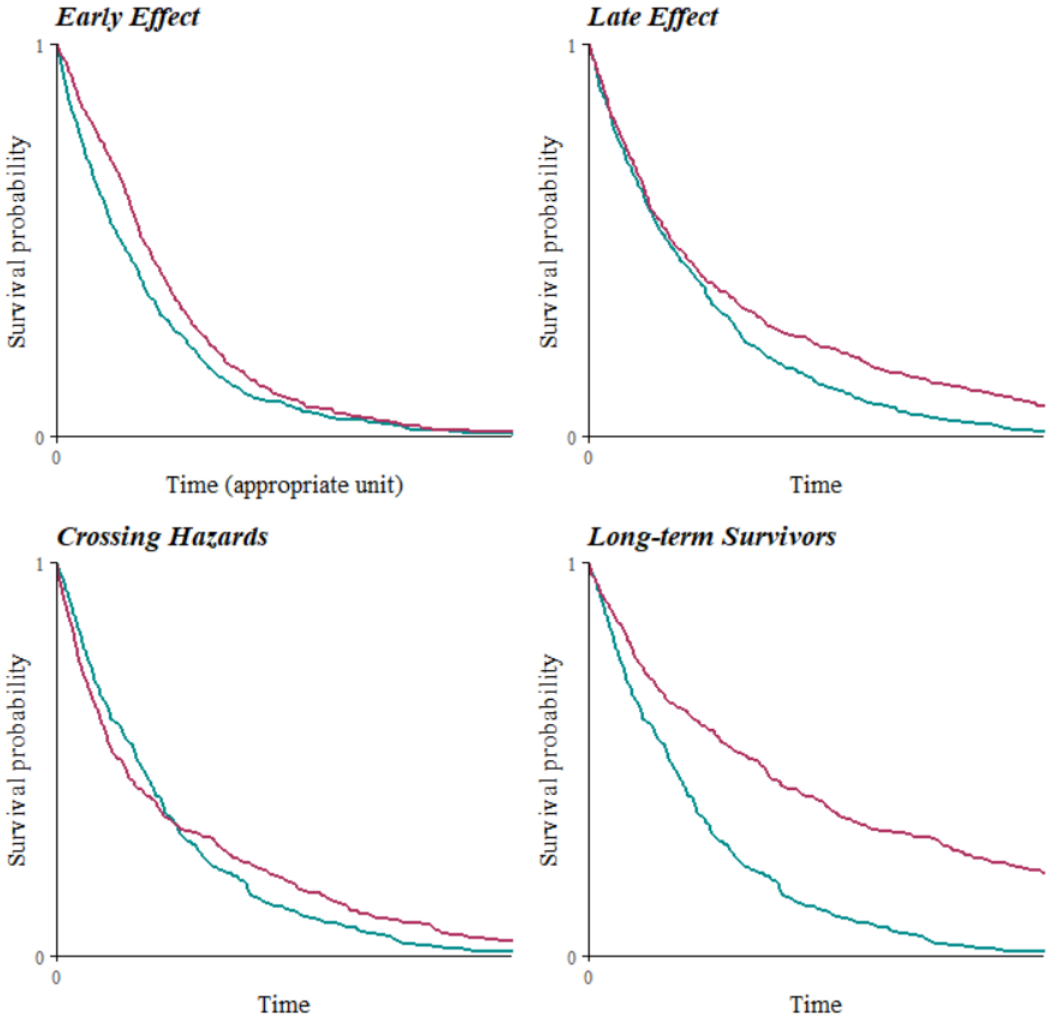

Figure 3.1: Patterns of non-proportionality.

3.2 Formal statistical tests

Since the introduction of the Cox model, its dependence on the proportionality assumption led to a series of statistical tests, with Cox (1972) being the first who suggested one via an extended version of the model. Almost a decade later, other statisticians started proposing tests for detecting different patterns of departure from proportionality, with many of them focusing on the alternative of monotonous hazard ratio or other specific time functions of the HR. Of course, many tests are considered omnibus, as they perform equally well for a wide range of alternatives. Notably, most of them employ similar techniques and ideas. For instance, the difference between observed and expected values arising from the Cox model is a repeatedly encountered quantity in tests for PH. As more and more methods were suggested, several generalizations, connecting comparable tests, were created (see for instance Grambsch and Therneau's general framework in section 3.2.5) setting the stage for more powerful testing procedures. At the same time, technology evolution allowed for the development of new approaches and the comparison of the existing ones through large simulation studies.

Even though it is difficult to define a clear classification since tests for detecting non-PH patterns oftentimes share similar characteristics, an attempt to do so can result in the following categorization:

1. Interval-dependent tests,
2. Tests based on weighting functions,
3. Score tests based on alternative models, and
4. Score process-based tests.

In the next sections, the most popular tests from each category, from 1972 till now, will be presented, and special attention will be paid to the two-sample case as the comparison of two groups is usually of great interest.

3.2.1 Interval-dependent tests

Interval-dependent tests require some arbitrary definition of time and/or covariate space partitions. The oldest of these tests was proposed by Schoenfeld (1980) and considers both time and covariate space partition. Two years later, Andersen (1982) brainstormed a rather innovative method to check the PH assumption, but it was still based on partitioning the space of the covariate whose proportionality is suspected.

Finally, the suggestion of Anderson & Senthilselvan (1982) and its generalizations by Moreau, O'Quigley & Mesbah (1985) and O'Quigley & Pessione (1989) combined the rationale behind the interval-dependent tests with proposed candidate forms of alternative models under the assumption that proportionality is invalid. Further insight on the aforementioned tests is given below.

Schoenfeld's test (1980)

Schoenfeld's proposal is an omnibus chi-square goodness of fit (GOF) test calculated for a proportional hazards model by obtaining the observed and expected numbers of deaths within each cell $C_{sj}, j = 1, 2, \ldots, r, s = 1, 2, \ldots, k$. Here a cell is defined as the combination of a time interval $[b_{j-1}, b_j)$ with a group A_s of particular characteristics (typically, $b_0 = 0$ and $b_r = \infty$). For instance, when two variables are under examination, e.g. gender (male or female) and smoking status (current smoker, former smoker, non-smoker), then it is reasonable to form six groups A_s, i.e.,

$A_1 \rightarrow$ men who are smokers,
$A_2 \rightarrow$ men who are former smokers,
$A_3 \rightarrow$ men who are non-smokers,
$A_4 \rightarrow$ women who are smokers,
$A_5 \rightarrow$ women who are former smokers, and,
$A_6 \rightarrow$ women who are non-smokers.

Each of these groups is further divided into follow-up periods according to the specified time intervals $[b_{j-1}, b_j), j = 1, 2, \ldots, r$ and the cell C_{sj} is subsequently created. The corresponding conditional mean of deaths e_{sj} and its variance v_{sj} in this cell, (and also covariances) are computed based on the partial likelihood arguments of Cox, given the risk set R_i at each failure time t_i of the *i-th* individual. Let D_j be the set of individuals who failed during the time interval $[b_{j-1}, b_j)$ and $\hat{\beta}$ the MPLE of the Cox model. Then,

$$\hat{e}_{sj} = \sum_{i \in D_j} \frac{\sum_{\ell \in R_i} I_s(\ell) \exp(\hat{\beta}' x_\ell)}{\sum_{\ell \in R_i} \exp(\hat{\beta}' x_\ell)}$$

and

$$\hat{v}_{sj} = \sum_{i \in D_j} \frac{\sum_{\ell \in R_i} I_s(\ell) \exp(\hat{\beta}' x_\ell)}{\sum_{\ell \in R_i} \exp(\hat{\beta}' x_\ell)} \left[1 - \frac{\sum_{\ell \in R_i} I_s(\ell) \exp(\hat{\beta}' x_\ell)}{\sum_{\ell \in R_i} \exp(\hat{\beta}' x_\ell)} \right]$$

where x_i is the covariate vector of the *i-th* subject and $I_s(.)$ is an indicator function with

$$I_s(\ell) = \begin{cases} 1, & \textit{if } \ell \in A_s \\ 0, & \textit{otherwise.} \end{cases}$$

In the formula of $\hat{e}_{sj}$, x_ℓ denotes all the covariates considered in the model, including the one whose proportionality is to be tested. After the calculation of the variance-covariance matrix V, and if the observed number of deaths is d_{sj} in the cell C_{sj}, then the suitable statistic for the PH assumption test is

$$Q = (d - \hat{e})' V^{-1} (d - \hat{e}) \tag{3.1}$$

where d is the vector of the observed number of events within each cell, and e is the vector of the expected number of events within each cell. Under the proportionality assumption, Q is asymptotically chi-square distributed with $(r-1) \times (k-1)$ degrees of freedom. For the two-sample case, the relative statistic is given by Moreau et al. (1985) below.

Clearly, the choice of partition here is of great importance. When the variables whose proportionality is tested are qualitative, it comes naturally to determine the partition of the covariate space based on their categories. For continuous variables, a common suggestion is to split their whole range into smaller intervals, usually in a meaningful way. Of course, depending on the sequence of these decisions, different results may occur. In any case, however, Schoenfeld (1980) stressed that if the partitions are defined so as to ensure that a similar number of events is contained in the cells, dissimilar choices will not result in substantially different outcomes. Finally, he suggested partitioning using all covariates involved in the PH model, even the ones whose proportionality is not in question (Song & Lee, 2000).

Andersen's test (1982)

In contrast to Schoenfeld's test, Andersen (1982) proposed partitioning of intervals based only on the covariate whose proportionality is suspected. Suppose that the PH assumption is not in question for the first p variables but may be invalid for the $(p+1)$-*th* covariate. Using the same notation as before, for a subject with label i, the covariate vector for the first p variables is $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$. To examine the proportionality of the $(p+1)$-*th* covariate, one must define k strata, according to the value of this variable, and r time intervals as before. Then, the s-*th* stratum has a hazard function

$$\lambda_{i,s}(t) = \lambda_{0s}(t) \exp(\beta' x_i), s = 1, 2, \ldots, k.$$

The baseline hazard function $\lambda_{0s}(t)$ in each stratum $s = 1, 2, \ldots, k$ can be approximated by a constant function λ_{sj} within each time interval $[b_{j-1}, b_j), j = 1, 2, \ldots, r$. These newly introduced $r \times k$ additional parameters can be computed using Breslow's maximum likelihood estimator on the condition that β is known. According to Andersen (1982), the following formula is used:

$$\hat{\lambda}_{sj} = \frac{d_{sj}}{\sum_{i=1}^{n_s} \exp(\beta' x_i^s) \cdot T_{isj}}, s = 1, 2, \ldots, k, j = 1, 2, \ldots, r,$$

where d_{sj} is the observed number of events during the *j-th* time interval in the *s-th* stratum, n_s is the number of subjects in the *s-th* group, x_i^s is the $p \times 1$ covariate vector for the *i-th* individual in statum s and finally, T_{isj} is the time spent in the *j-th* interval by the *i-th* individual from group s. Of course, when there is only one variable whose proportionality is going to be tested, the term $\exp(\beta' x_i^s)$ is removed from the previous formula, leaving only the total time spent in the *j-th* interval by the subjects in group s in the denominator.

Under the PH assumption, it is expected that $\lambda_{(s+1)j} = \lambda_{sj} \exp(a_{s+1}), \forall s, j$. Retrospectively, this is equivalent to $\ln \lambda_{(s+1)j} = \ln \lambda_{1j} + \sum_{l=2}^{s+1} a_l, \forall s, j$. If $\ln \lambda_{1j}$ is denoted by ξ_j, then

$$\hat{a}_{s+1} = \frac{\sum_{j=1}^{r} [\ln \hat{\lambda}_{(s+1)j} - \ln \hat{\lambda}_{sj}] / [d_{(s+1)j}^{-1} + d_{sj}^{-1}]}{\sum_{j=1}^{r} [d_{(s+1)j}^{-1} + d_{sj}^{-1}]^{-1}}, s = 1, 2, \ldots, k-1,$$

which is a weighted average of the difference $\ln \lambda_{(s+1)j} - \ln \lambda_{sj}$ with weights depending on the number of deaths in the s-th stratum, and

$$\hat{\xi}_j = \frac{\sum_{s=1}^{k} d_{sj} (\ln \hat{\lambda}_{sj} - \sum_{l=1}^{s} \hat{a}_l)}{\sum_{s=1}^{k} d_{sj}}, j = 1, 2, \ldots, r,$$

where $\hat{a}_1 \equiv 0$.

Subsequently, the PH assumption can be tested via the statistic[3]

$$Q = \sum_{j=1}^{r} \sum_{s=1}^{k} d_{sj} [\ln \hat{\lambda}_{sj} - (\hat{\xi}_j + \sum_{l=1}^{s} \hat{a}_l)]^2, \tag{3.2}$$

which asymptotically follows a chi-square distribution with $(r-1) \times (k-1)$ degrees of freedom. Small values indicate that $\ln \hat{\lambda}_{sj}$ and $\hat{\xi}_j + \sum_l \hat{a}_l$ are in general close. On the contrary, the PH assumption is rejected when significant differences between these two quantities are observed. Once again, the choice of partition plays a major

[3]Andersen also proposed another statistic for the PH assumption test which is proved to be asymptotically equivalent to the one presented here.

role in the outcome, but Andersen has pointed out that different partitions give the same results if each interval contains a reasonable number of failures.

Despite Andersen's ingenuity, many claim that this test is not suitable for the assessment of the validity of the PH assumption. When approximating the baseline hazard within each stratum and time interval via a constant function, Andersen initiates a GOF testing procedure which does not examine the fit of the Cox model. Nevertheless, others embrace his approach because they believe these two models are equivalent in practice.

Anderson and Senthilselvan's model (1982)

The violation of the proportionality of hazards led to the proposal of a new model which assumes that coefficients are different amongst non-overlapping time intervals. When only two time intervals are considered, the two-step model (Anderson & Senthilselvan, 1982), as they called it, has the following form:

$$\lambda_i(t) = \begin{cases} \lambda_0(t)\exp(\alpha' x_i), & \text{if } t < b_1 \\ \lambda_1(t)\exp(\gamma' x_i), & \text{if } t \geq b_1. \end{cases}$$

This two-step model can also be regarded as a model with constant coefficients but time-varying covariates. Here, α, γ and b_1 should be estimated along with the baseline hazard function. The typical procedure is to start by fixing b_1 and use the conditional likelihood to compute α and γ. Ideally, b_1 should be chosen in a way that ensures that enough events happen in the second time interval. If this condition is not met, γ will be poorly estimated and infinite estimates of some of its elements can arise with binary covariates (Senthilselvan, 1980). Finally, the baseline hazard function is estimated conditional on the estimates $\hat{\alpha}, \hat{\gamma}$ and b_1, using a penalized maximum likelihood[4].

Anderson and Senthilselvan's model can be generalized for more than two intervals. Nevertheless, the authors pointed out the difficulties of extending their method to several intervals, as the simultaneous estimation of several parameters is computationally demanding and the introduction of further censoring results in increased imprecision. Despite this issue, this model paved the way for a novel approach to survival data analysis and the creation of two PH assumption tests, which will be discussed below.

[4]Method proposed by Good & Gaskins (1971). It provides a smooth estimate of $\lambda_0(t)$.

Moreau, O'Quigley & Mesbah (1985)

Based on Anderson and Senthilselvan's model (1982), Moreau, O'Quigley, and Mesbah (1985) brainstormed the idea of testing the PH assumption by performing a score test on the coefficients of an alternative model. Quite similar to the previous one, this new model has the form

$$\lambda_i(t) = \lambda_0(t) \exp\{(\beta + \gamma_j)' x_i\}, \tag{3.3}$$

where β and γ_j are $r+1$, $p \times 1$ vectors and $t \in [b_{j-1}, b_j), j = 2, ..., r$. This means that depending on the time interval of interest the coefficients differ. Of course, without loss of generality, γ_1 can be assumed equal to zero, since there are only r intervals but $r+1$ coefficients of unknown parameters. Therefore,

$$\lambda_i(t) = \begin{cases} \lambda_0(t) \exp\{\beta' x_i\}, & \text{if } t \in [0, b_1) \\ \lambda_0(t) \exp\{(\beta + \gamma_2)' x_i\}, & \text{if } t \in [b_1, b_2) \\ \vdots & \\ \lambda_0(t) \exp\{(\beta + \gamma_r)' x_i\}, & \text{if } t \in [b_{r-1}, \infty). \end{cases}$$

After simplifying the formula a little, the null hypothesis of the test for proportionality is

$$H_0 : \gamma_2 = \ldots = \gamma_r = 0$$

versus the alternative

$$H_1 : \gamma_j \neq 0 \text{ for at least one } j \in \{2, \ldots, r\}.$$

A score test can be performed to test this hypothesis. The first and the second derivatives of the partial log-likelihood are needed since a score test statistic is given by the formula

$$S = U' I^{-1} U,$$

where U is the $rp \times 1$ vector of first derivatives and I is the $rp \times rp$ Fisher information matrix, i.e., the negative of the matrix of the second derivatives, calculated under the null hypothesis, using the MPLE of the simple Cox model $\hat{\beta}$ as β and $\gamma_j = 0$, for $j = 2, \ldots, r$.

According to (2.8), and assuming that $t_{ij}, i = 1, 2, \ldots, k_j, j = 1, 2, \ldots, r$ are the distinct survival times in the j-th interval, the partial log-likelihood for (3.3) is

$$\ell(\gamma_2, \ldots, \gamma_r, \beta) = \sum_{j=1}^{r} \sum_{i=1}^{k_j} \left[(\beta + \gamma_j)' x_i - \ln \sum_{\ell \in R_{ij}} \exp\{(\beta + \gamma_j)' x_\ell\} \right] \tag{3.4}$$

where R_{ij} is the risk set at time t_{ij}. The $rp \times 1$ vector of first derivatives U can be split into r parts of $p \times 1$ vectors U_j, where each one has p values V_{sj}. V_{sj} is the first derivative of (3.4) with respect to $\gamma_{sj}, s = 1, \ldots, p, j = 2, \ldots, r$ calculated under H_0:

$$V_{sj} = \sum_{i=1}^{k_j} (x_{is} - A_{isj})$$

where

$$A_{isj} = \frac{\sum_{\ell \in R_{ij}} x_{\ell s} \exp(\hat{\beta}' x_\ell)}{\sum_{\ell \in R_{ij}} \exp(\hat{\beta}' x_\ell)}.$$

So,

$$U_j' = (V_{1j}, \cdots, V_{pj}), \text{ for } j = 2, \ldots, r,$$

and U can be written as

$$U' = (U_2', \cdots, U_r', 0)$$

due to the fact that the p elements corresponding to $\partial \ell / \partial \beta_s$ are equal to zero when $\beta = \hat{\beta}$. Notice that U_1 is not defined, but a natural way to do so is by replacing j with 1 in the formula of V_{sj}. Then, interestingly,

$$\sum_{j=1}^{r} U_j' = \left(\frac{\partial \ell}{\partial \beta_1}(0, \cdots, 0, \hat{\beta}), \cdots, \frac{\partial \ell}{\partial \beta_r}(0, \cdots, 0, \hat{\beta}) \right) = 0. \tag{3.5}$$

Regarding the second derivatives, it holds that under H_0,

$$-\frac{\partial^2 \ell}{\partial \gamma_{sj} \partial \gamma_{qj}} = -\frac{\partial^2 \ell}{\partial \beta_s \partial \gamma_{qj}} = \sum_{i=1}^{k_j} \left[\frac{\sum_{\ell \in R_{ij}} x_{\ell s} x_{\ell q} \exp(\hat{\beta}' x_\ell)}{\sum_{\ell \in R_{ij}} \exp(\hat{\beta}' x_\ell)} - A_{isj} A_{iqj} \right] \tag{3.6}$$

for $j = 2, \ldots, r$, and $s, q \in \{1, 2, \ldots, p\}$, and

$$-\frac{\partial^2 \ell}{\partial \gamma_{sj} \partial \gamma_{qj'}} = 0$$

when $j \neq j'$.

Let $I_j, j = 2, \ldots, r$ be a $p \times p$ matrix with elements the second derivatives corresponding to (3.6). Again, I_1 is not defined but it can occur by replacing j with 1 in the same equation. Then, it can be shown that

$$-\frac{\partial^2 \ell}{\partial \beta_s \partial \beta_q} = \sum_{j=1}^{r} I_j.$$

Having defined all the above formulas, the form of the observed information matrix I can be easily written as

$$\begin{pmatrix} D & B \\ B' & C \end{pmatrix}$$

where $D = \text{diag}(I_2, \cdots, I_r)$, $B' = (I_2, \cdots, I_r)$ and $C = \sum_{j=1}^{r} I_j$. Fortunately, after implementing Rao's method (Rao et al., 1973), this matrix is inverted, and after calculations, it occurs that the score statistic for the test of proportionality is

$$S = \sum_{j=1}^{r} U_j' I_j^{-1} U_j. \tag{3.7}$$

Since the null hypothesis assumes that $(r-1)p$ values are equal to zero, the asymptotic distribution of S is a chi-square with $(r-1)p$ degrees of freedom. Notice that the calculation of S requires only the inversion of I_j, making the procedure simpler than it seemed initially. As for the occasion where more than one death may occur in a single time point t_{ij}, i.e., when there are ties, small adjustments should be made in the formulas of first and second derivatives of the partial log-likelihood. More specifically, if d_{ij} is the number of deaths at t_{ij},

$$V_{sj} = \sum_{i=1}^{k_j} (x_{is} - d_{ij} A_{isj}),$$

and

$$-\frac{\partial^2 \ell}{\partial \gamma_{sj} \partial \gamma_{qj}} = -\frac{\partial^2 \ell}{\partial \beta_s \partial \gamma_{qj}} = \sum_{i=1}^{k_j} d_{ij} \left[\frac{\sum_{\ell \in R_{ij}} x_{\ell s} x_{\ell q} \exp(\hat{\beta}' x_\ell)}{\sum_{\ell \in R_{ij}} \exp(\hat{\beta}' x_\ell)} - A_{isj} A_{iqj} \right].$$

To gain more insight into Moreau, O'Quigley, and Mesbah's proposed test, the two-sample case will be examined theoretically and via simulations (see Chapter 4 for more). First of all, when the model includes one single variable, the score statistic has the following form:

$$S = \sum_{j=1}^{r} \frac{U_j^2}{I_j}, \tag{3.8}$$

since $p = 1$. Suppose two groups and thus, a variable x with two possible values, with the usual notation being

$$x = \begin{cases} 1, & \text{if the subject belongs to the } 1^{st} \text{ group} \\ 0, & \text{otherwise.} \end{cases}$$

In the absence of tied data and according to what was previously presented, U_j is just the difference between observed and expected number of deaths in group 1, i.e.,

$$U_j = O_{1j} - E_{1j}, \tag{3.9}$$

where

$$E_{1j} = \sum_{i=1}^{k_j} e_{ij}, \text{ with } e_{ij} = \frac{r_{1ij}e^{\hat{\beta}}}{r_{1ij}e^{\hat{\beta}} + r_{2ij}}, \ j = 1, 2, \ldots, r.$$

The quantities r_{1ij} and r_{2ij} represent the number of individuals at risk at time t_{ij} in group 1 and group 2, respectively. Moreover,

$$I_j = \sum_{i=1}^{k_j} e_{ij}(1 - e_{ij}). \tag{3.10}$$

Replacing U_j's and I_j's with their equivalent quantities from (3.9) and (3.10) in (3.8), the score statistic for the two-sample case is ready to be used. Here, it follows a chi-square distribution with $r - 1$ degrees of freedom. Notice that it is identical to the one given by Schoenfeld (1980). For that reason, the same paper suggested a more conservative version that takes the form of a typical chi-square test, i.e.,

$$\sum_{j=1}^{r} \left[\frac{(O_{1j} - E_{1j})^2}{E_{1j}} + \frac{(O_{2j} - E_{2j})^2}{E_{2j}} \right].$$

Of course, the expected number of deaths in group 2 is

$$E_{2j} = \sum_{i=1}^{k_j} (1 - e_{ij}), \ j = 1, 2, \ldots, r,$$

because the total number of deaths in the j-th interval is constant and $E_{1j} + E_{2j} = k_j, \forall j \in \{1, 2, \cdots, r\}$. The reason why this statistic is more conservative than the one presented before stems from the Cauchy-Schwarz inequality: it holds that

$$\sum_{i=1}^{k_j} e_{ij}^2 \geq \frac{\left(\sum_{i=1}^{k_j} e_{ij}\right)^2}{k_j}.$$

Consequently, $\forall j \in \{1, 2, \ldots, r\}$

$$\frac{(O_{1j} - E_{1j})^2}{E_{1j}} + \frac{(O_{2j} - E_{2j})^2}{E_{2j}} = \frac{k_j(O_{1j} - E_{1j})^2}{E_{1j}E_{2j}} = \frac{(O_{1j} - E_{1j})^2}{\sum e_{ij} - \frac{(\sum e_{ij})^2}{k_j}} \leq \frac{U_j^2}{I_j}.$$

Finally, it is important to stress that the aforementioned statistics must be modified in the presence of tied data. Once again, the number of deaths at each time point t_{ij} is required and the formulas of E_{1j}, E_{2j} and I_j change as follows:

$$E_{1j} = \sum_{i=1}^{k_j} d_{ij}e_{ij},$$

$$E_{2j} = \sum_{i=1}^{k_j} d_{ij}(1 - e_{ij}),$$

and

$$I_j = \sum_{i=1}^{k_j} d_{ij} e_{ij}(1 - e_{ij}).$$

O'Quigley & Pessione's test (1989)

Introducing a time axis division, O'Quigley and Pessione (1989) suggested a test for the PH assumption based on the model

$$\lambda_i(t) = \lambda_0(t) \exp\{(\beta + \Psi_j \theta)' x_i\}, \ t \in [b_{j-1}, b_j), \ j = 1, 2, \ldots, r, \tag{3.11}$$

where β and θ are $p \times 1$ unknown vectors and Ψ_j is a $p \times p$ diagonal matrix with diagonal elements $\psi_{1j}, \psi_{2j}, \ldots, \psi_{pj}$. When $\theta = 0$, the model in (3.11) reduces to the simple Cox PH model. Therefore, the null hypothesis

$$H_0 : \theta = 0,$$

versus the alternative

$$H_A : \theta \neq 0$$

may be tested by using a score statistic. Its formula is derived from the first and second derivatives of the partial log-likelihood of the model in (3.11), in a similar manner to the previous test by Moreau et al. (1985). For the two-sample case, which will be implemented in Chapter 4, the score statistic S is equal to U^2/I, where

$$U = \sum_{j=1}^{r} \Psi_j (O_{1j} - E_{1j}) = \sum_{j=1}^{r} \Psi_j (O_{1j} - \sum_{i=1}^{k_j} e_{ij})$$

and

$$I = \sum_{j=1}^{r} \sum_{i=1}^{k_j} \Psi_j^2 e_{ij}(1 - e_{ij}) - \frac{\left[\sum_{j=1}^{r} \sum_{i=1}^{k_j} \Psi_j e_{ij}(1 - e_{ij})\right]^2}{\sum_{j=1}^{r} \sum_{i=1}^{k_j} e_{ij}(1 - e_{ij})}.$$

Here, O_{1j}, E_{1j} and e_{ij} are defined as before. As for the scalar[5] values Ψ_j, these are chosen by the researcher performing the test. Some suggestions from O'Quigley and Pessione (1989) for the general case of p covariates are:

[5]Here $p = 1$, and thus Ψ_j is considered to be a real number, not a matrix.

1. *Linear trend alternative*:

$$\psi_{qj} = j - 1, \text{ for } j = 1, 2, \ldots, r$$

$$\psi_{q'j} = 0, \text{ for } q' \neq q \text{ and } j = 1, 2, \ldots, r$$

2. *Exponential decay*:

$$\psi_{qj} = r^{-1} + \cdots + (r - j + 1)^{-1}, \text{ for } j = 1, 2, \ldots, r$$

$$\psi_{q'j} = 0, \text{ for } q' \neq q \text{ and } j = 1, 2, \ldots, r$$

3. *Inversion of regression effect for the two-sample case with two time intervals*:

   When inversion of effect is suspected, a proposal is to use $\Psi_1 = 1$ and $\Psi_2 = -1$. This suggestion is suitable for the case of crossing hazards. In general, one may choose any pair of Ψ_1 and Ψ_2 for which $\Psi_1 = -\Psi_2$. If effects do go in opposite directions but are not of comparable magnitude, then the formulation used here may not be very efficient. However, if the differences in magnitude are large enough, even a model with an inappropriate assumption such as proportional hazards will detect differential effects.

The choice of the matrix Ψ_j is important since the interpretation of a non-zero value for θ will depend on it.

Last but not least, one should keep in mind that the statistic given above is only appropriate when there are no ties. Once again, a modification is needed if that is not the case. For the two-sample problem, U and I become

$$U = \sum_{j=1}^{r} \Psi_j (O_{1j} - \sum_{i=1}^{k_j} d_{ij} e_{ij})$$

and

$$I = \sum_{j=1}^{r} \sum_{i=1}^{k_j} d_{ij} \Psi_j^2 e_{ij}(1 - e_{ij}) - \frac{\left[\sum_{j=1}^{r} \sum_{i=1}^{k_j} d_{ij} \Psi_j e_{ij}(1 - e_{ij})\right]^2}{\sum_{j=1}^{r} \sum_{i=1}^{k_j} d_{ij} e_{ij}(1 - e_{ij})}.$$

3.2.2 Tests based on weighting functions

This group of tests employs weighting functions, particularly from Fleming & Harrington's original or extended family, and has been proved to be powerful when the alternative hypothesis is that of increasing or decreasing (monotonous) hazard ratio over time. In this category, three tests are worthwhile to mention: Gill & Schumacher's (1987), Lin's (1991), and Sengupta, Bhattacharjee & Rajeev's (1998).

Gill and Schumacher's test (1987)

This test is based on the comparison of different generalized rank estimators of the proportionality constant θ and in its original form, it is appropriate only for the two-sample problem, i.e., when the interest is focused on two subgroups of the data. To begin with, let $N_j(t)$ denote the number of failures in group j before or at t, $Y_j(t)$ the number at risk in group j at t, $\lambda_j(t)$ the hazard and $\Lambda_j(t)$ the cumulative hazard rate at t, for $j = 1, 2$. Here the test problem is given by H_0 versus H_A where

$$H_0 : \frac{\lambda_2(t)}{\lambda_1(t)} = \theta, \text{ for some positive number } \theta,$$

and

$$H_A : \frac{\lambda_2(t)}{\lambda_1(t)} \neq \theta, \text{ for any positive number } \theta.$$

Under H_0, θ can be estimated by the generalized rank estimator

$$\theta_K = \frac{\int_0^\tau K(t) d\Lambda_2(t)}{\int_0^\tau K(t) d\Lambda_1(t)},$$

where τ is the upper limit of observable survival times, and $K(t)$ is a weighting function, typically chosen from the Fleming-Harrington (FH) family, the initial version of which is given by

$$K_{FH}(t) = \frac{Y_1(t)Y_2(t)}{Y_1(t) + Y_2(t)} \{S(t)\}^\rho, \rho > 0. \tag{3.12}$$

So, in practice, in order to estimate θ, one needs to compute $Y_j(t)$, $S(t)$, and $\Lambda_j(t)$. Gill & Schumacher (1987) proposed to estimate $S(t)$ using the right continuous version of the Kaplan–Meier estimator and $\Lambda_j(t), j = 1, 2$, implementing Nelson–Aalen's approach (see sections 2.2 and 2.4.5). Under H_0, θ_K converges in probability to θ as the sample size increases. Consequently, for large sample sizes, the difference between θ_{K_1} and θ_{K_2} for two different weight functions is expected to be small. On the other hand, when H_A holds, one anticipates gaining quite dissimilar estimates of the hazard ratio, since two weight functions will yield estimates emphasizing on and representing better different follow-up periods. Having said all that, it is reasonable to base the test on the difference between two rank estimators of the aforementioned form.

Let $K_1(t)$ and $K_2(t)$ be two weighting functions of the form presented in (3.12), i.e.,

$$K_i(t) = \frac{Y_1(t)Y_2(t)}{Y_1(t) + Y_2(t)} \{S(t)\}^{\rho_i}$$

for $i \in \{1, 2\}$, with $\rho_1 \neq \rho_2$. Then for the j-th group and the i-th weighting function, the quantities $\hat{K}_{ij}$ are defined as

$$\hat{K}_{ij}(t) = \int_0^\tau \hat{K}_i(t) d\hat{\Lambda}_j(t), i, j \in \{1, 2\}$$

and thus,

$$\hat{\theta}_{K_i}(t) = \frac{\hat{K}_{i2}}{\hat{K}_{i1}}, i = 1, 2.$$

Instead of using the difference $\hat{\theta}_{K_2}(t) - \hat{\theta}_{K_1}(t) = \hat{K}_{22}/\hat{K}_{21} - \hat{K}_{12}/\hat{K}_{11}$, a symmetrized version is considered as a test statistic:

$$Q_{K_1K_2} = \hat{K}_{22}\hat{K}_{11} - \hat{K}_{21}\hat{K}_{12}$$

The asymptotic variance of $Q_{K_1K_2}$ can be estimated by

$$\hat{S}^2_{Q_{K_1K_2}} = \hat{K}_{21}\hat{K}_{22}\hat{V}_{11} - \hat{K}_{21}\hat{K}_{12}\hat{V}_{12} - \hat{K}_{11}\hat{K}_{22}\hat{V}_{21} + \hat{K}_{11}\hat{K}_{12}\hat{V}_{22}$$

where

$$V_{ii'} = \int_0^\tau \frac{K_i(t)K_{i'}(t)}{Y_1(t)Y_2(t)} d\{N_1(t) + N_2(t)\}.$$

Of course, when K_i and $K_{i'}$ are chosen from the FH family, the denominator inside the interval is erased. Since the standard procedure involves employing the log-rank and Prentice's Wilcoxon weight function, which are given by replacing ρ with 0 and 1 in (3.12), respectively, this is usually the case. In general, a fascinating fact is that if the ratio $\frac{K_2(t)}{K_1(t)}$ is monotonous, then the test achieves maximum power under alternatives with a monotone hazard ratio. The latter is always fulfilled when FH weights are selected.

That being said, the following standardized statistic can be used for the implementation of the test for proportional hazards:

$$T_{K_1K_2} = \frac{Q_{K_1K_2}}{\hat{S}_{Q_{K_1K_2}}}. \tag{3.13}$$

Under H_0 and as the sample size increases, $T_{K_1K_2}$ has a standard normal distribution.

Despite its simplicity, Gill and Schumacher's test has two major disadvantages. The first is that the variance estimator of $Q_{K_1K_2}$ may be negative, especially far from the null hypothesis. The second is that the test cannot be implemented for continuous covariates or qualitative variables with more than two categories. Nevertheless, some comments were made by the authors in the relative paper regarding the latter occasion: when there are k groups in the data, a global test can be performed

combining all possible pairwise comparisons. The resulting statistic will then have, asymptotically, and under the null hypothesis, a chi-square distribution with $k-1$ degrees of freedom. While feasible, such a procedure is usually avoided due to its computational complexity and other tests are preferred for this cause.

Lin's test (1991)

Lin's (1991) proposal has a similar spirit to the previous test, but it allows for simultaneous testing of the PH assumption for many covariates. It is based on the difference between the Cox PH model's MPLE $\hat{\beta}$ and a weighted counterpart $\hat{\beta}_w$. The latter occurs when a weighting function $w(t)$ is introduced into the partial likelihood score equation, i.e.,

$$U_w(\beta) = \sum_{i=1}^{n} \delta_i w(t_i) \left[x_i - \frac{\sum_{\ell \in R_i} x_\ell \exp(\beta' x_\ell)}{\sum_{\ell \in R_i} \exp(\beta' x_\ell)} \right],$$

where δ_i is the event indicator, x_i is the covariate vector, R_i is the risk set for subject i and finally, $w(t_i)$ is a weighting function evaluated over $t = t_i$, i.e., the time of event or censoring for an individual with label $i \in \{1, 2, \ldots, n\}$.

After solving the equation $U_w(\beta) = 0, \hat{\beta}_w$ occurs, a $p \times 1$ vector which should be close to $\hat{\beta}$ under the null hypothesis of proportionality. With that in mind, Lin suggested performing the test utilizing the statistic

$$Q_w = n(\hat{\beta}_w - \hat{\beta})'[C_w(\hat{\beta}) - C(\hat{\beta})]^{-1}(\hat{\beta}_w - \hat{\beta})$$

which asymptotically follows a chi-square distribution with p degrees of freedom. The quantities $C_w(\hat{\beta}), C(\hat{\beta})$ are the variance-covariance matrices of $n^{1/2}(\hat{\beta}_w - \beta_0)$ and $n^{1/2}(\hat{\beta} - \beta_0)$, respectively, under the null hypothesis and the assumption that the real hazard ratio is equal to a constant value β_0. Therefore, one can replace the $C_w(\hat{\beta}) - C(\hat{\beta})$ with the difference between $\hat{\beta}_w$'s and $\hat{\beta}$'s covariance matrices multiplied by n.

In the two-sample case, the test does not reduce to Gill and Schumacher's (1987) test for proportionality, despite their similarity. Also, the variance estimator here is always non-negative.

Sengupta, Bhattacharjee & Rajeev's test (1998)

Sengupta et al. (1998) proposed a two-sample test against the alternative of increasing **cumulative** hazard ratio. More specifically, here,

$$H_0 : \frac{\Lambda_2(t)}{\Lambda_1(t)} = \theta \text{ for some positive number } \theta,$$

versus

$$H_A : \frac{\Lambda_2(t)}{\Lambda_1(t)} \text{ is an increasing function of time.}$$

It is easy to see that if the cumulative hazard rates of two groups are proportional, the same holds for their hazard functions and vice versa, but a monotonous hazard ratio is a special case of a monotonous cumulative hazard ratio. So, in a sense, this is a generalization of the test proposed by Gill & Schumacher (1987).

Again, let $N_j(t)$ denote the number of failures in group j before or at t, $Y_j(t)$ the number at risk in group j at t, and $\Lambda_j(t)$ the cumulative hazard rate for $j = 1, 2$. Now, define K_{ij} as follows:

$$K_{ij} = \int_0^\tau K_i(t)\Lambda_j(t)dt, \text{ with } i, j \in \{1, 2\}.$$

Then,

$$Q_{K_1K_2} = \hat{K}_{22}\hat{K}_{11} - \hat{K}_{21}\hat{K}_{12}$$

and its estimated variance is given by the formula

$$\hat{S}^2_{Q_{K_1K_2}} = \hat{K}_{21}\hat{K}_{22}\hat{V}_{11} - \hat{K}_{21}\hat{K}_{12}\hat{V}_{12} - \hat{K}_{11}\hat{K}_{22}\hat{V}_{21} + \hat{K}_{11}\hat{K}_{12}\hat{V}_{22}$$

where

$$V_{ii'} = \int_0^\tau \int_0^\tau K_i(t)K_{i'}(s)V(s \wedge t)dsdt$$

and

$$V(t) = \int_0^t \frac{dN_1(s) + dN_2(s)}{Y_1(s)Y_2(s)}.$$

As expected, the statistic

$$T_{K_1K_2} = \frac{Q_{K_1K_2}}{\hat{S}_{Q_{K_1K_2}}} \tag{3.14}$$

asymptotically follows a standard normal distribution. However, careful consideration should be given to the fact that, according to the relative paper, $Q_{K_1K_2}$ is zero under H_0 and positive under H_A. Consequently, it is preferable to use the normalized statistic $T_{K_1K_2}$ to perform a one-sided test, in contrast to the previous cases where two-sided tests are suggested.

3.2.3 Score tests based on alternative models

This is a rather broad category. Some of the tests already described are score tests (see for example: Moreau et al., 1985; O'Quigley & Pessione, 1989). Nevertheless, other characteristics seem to be dominant and definitive for their classification (e.g.

their dependence on the choice of time intervals). Here, methods suggested by Breslow et al. (1984), Quantin et al. (1996), Bagdonavičius et al. (2004), Bagdonavičius & Levulienė (2019), Kraus (2007), and Hafdi (2021) are briefly presented.

Breslow et al. (1984)

Breslow, Elder and Berger's (1984) proposal, also known as the *acceleration test*, is a two-sample testing procedure for the assumption of proportional hazards. Its rationale is inspired by the extended Cox model, according to which, one or more covariates may be time dependent. To be more specific, the extended Cox model is usually presented in the form

$$\lambda_i(t) = \lambda_0(t) \exp(\beta' x_i + \gamma' z_i(t))$$

where x_i is a $p \times 1$ vector of fixed covariates and $z_i(t)$ is a $q \times 1$ vector of time-dependent variables. For the two-sample case, i.e., when there is only one dichotomous variable taking the value 1 for subjects who belong to the first group and 0 for those who belong to the second, the aforementioned equation can be written as

$$\lambda_1(t) = \lambda_0(t) \exp(\beta + \gamma z(t)). \tag{3.15}$$

Of course, $\lambda_1(t)$ corresponds to the hazard function of a subject in the first group, while the hazard for the second group is the baseline function $\lambda_0(t)$. Breslow et al. (1984) suggested testing the PH assumption via a score test on the previous model. Thus, one must test the null hypothesis

$$H_0 : \gamma = 0,$$

versus the alternative

$$H_A : \gamma \neq 0.$$

The procedure for the calculation of the test statistic here is simple (see for instance Moreau et al. (1985) above), and thus the related steps will be omitted. Let $t_j, j = 1, 2, \ldots, m$ denote the m distinct ordered failure times (there are no ties), $d_{1j} = 1–d_{2j}$ an indicator variable which is equal to 1 if the event at t_j occurred for an individual in the first group, and finally, r_{1j} and r_{2j} the number of subjects at risk at t_j in group 1 and group 2, respectively. Then, the conditional probability that the event at t_j is from sample 1, is given by

$$p_j = p_j(\beta, \gamma) = \frac{\exp\{\beta + \gamma z(t_j)\} r_{1j}}{\exp\{\beta + \gamma z(t_j)\} r_{1j} + r_{2j}} \tag{3.16}$$

and therefore, $q_j = 1-p_j$ is the conditional probability that the failure at t_j happened to a subject from sample 2. Under H_0, $\gamma = 0$ and the conditional probabilities for the two groups are

$$\hat{p}_j = p_j(\hat{\beta}, 0) = \frac{e^{\hat{\beta}} r_{1j}}{e^{\hat{\beta}} r_{1j} + r_{2j}}$$

and

$$\hat{q}_j = 1\text{–}\hat{p}_j = \frac{r_{2j}}{e^{\hat{\beta}} r_{1j} + r_{2j}}$$

where $\hat{\beta}$ is the MPLE under the null hypothesis. According to Breslow et al. (1984), based on the partial likelihood of the model in (3.15), the score test statistic is

$$\frac{U^2}{I} \tag{3.17}$$

where

$$U = \sum_{j=1}^{m} z_j (d_{1j} - \hat{p}_j)$$

and

$$I = \sum_{j=1}^{m} z_j^2 \hat{p}_j \hat{q}_j - \frac{\left[\sum_{j=1}^{m} z_j \hat{p}_j \hat{q}_j\right]^2}{\sum_{j=1}^{m} \hat{p}_j \hat{q}_j}.$$

As expected, under H_0 it asymptotically follows a χ^2 distribution with 1 degree of freedom. While this is exactly the test that Cox (1972) proposed, the authors managed to take his idea one step forward. Notice that in the final formula of the statistic, the quantities $z_j, j = 1, 2, \ldots, m$ play a major role. A well-known practice, originally suggested by Cox, is to use the failure times t_j or their logarithms as z_j. However, this choice will cause the acceleration test to fail to be invariant under monotone transformations of the survival times. A better approach will be to use the rank scores $z_j = j$ or the cumulative hazard scores $z_j = \sum_{l=1}^{j} 1/r_l$ (Nelson–Aalen estimator of the cumulative hazard when there are no ties). Both sets have the desired feature that they are monotone increasing in j and depend only on rank information. Of course, other choices are possible but these two seem to be quite powerful, especially in the case of crossing hazards. Nevertheless, every method has its drawbacks. For instance, the rank score test is heavily influenced by the censoring distribution, while the cumulative hazards scores are not a good choice when the sample size is small. Further work is needed to determine the optimal assignment of values to the z_j.

Once more, an alteration is required when more than one events may happen at the same time point. More specifically, U and I from (3.17) become

$$U = \sum_{j=1}^{m} z_j (d_{1j} - \mu_j(\hat{a}))$$

and

$$I = \sum_{j=1}^{m} z_j^2 \sigma_j^2(\hat{a}) - \frac{\left[\sum_{j=1}^{m} z_j \sigma_j^2(\hat{a})\right]^2}{\sum_{j=1}^{m} \sigma_j^2(\hat{a})},$$

where $\mu_j(\hat{a})$ and $\sigma_j^2(\hat{a})$ denote the mean and the variance of the noncentral (Fisher's) hypergeometric distribution which, the number of failures in group 1, d_{1j} follows under H_0, given the total number of deaths d_j at t_j, and the number of individuals at risk r_{1j} and r_{2j} in groups 1 and 2, respectively. Here, $\hat{a}$ is determined as the solution of the equation $\sum_j d_{1j} = \sum_j \mu_j(a)$. Due to the complexity of this procedure, an easier approach is to substitute a binomial distribution for the hypergeometric one. The quantities $d_j p_j(\hat{a})$ and $d_j p_j(\hat{a}) q_j(\hat{a})$ replace $\mu_j(\hat{a})$ and $\sigma_j^2(\hat{a})$, respectively, both for determination of $\hat{a}$ and in the calculation of the test statistic. This approximate version is quite accurate when most of the d_j are small in comparison to the corresponding numbers of subjects at risk in each group. Since this is the most common scenario, and also, due to the fact that the approximate statistic agrees precisely with the original version when $d_j = 1$, it is preferable. As for the value $\hat{a}$ another approximation can be used in order to avoid the iterative calculations: Breslow et al. (1984) suggested to calculate $\hat{a}$ from the Mantel–Haenszel estimator. Last but not least, it is important to stress that in the presence of ties the cumulative hazard scores are also somewhat different, and they are given by the formula

$$z_j = \sum_{\ell=1}^{j} \frac{d_\ell}{r_\ell}.$$

Quantin et al. (1996)

The approach that is to be presented in this section is simply a generalization of the previous test. The proposed model in terms of cumulative hazards is

$$\Lambda_i(t) = \exp(\beta' x_i)\{\Lambda_0(t)\}^{\exp(\gamma' x_i)},$$

which means that the usual presentation in terms of hazard function takes the form of

$$\lambda_i(t) = \lambda_0(t) \exp[\beta' x_i + \gamma' x_i + \{\exp(\gamma' x_i) - 1\} \ln \Lambda_0(t)]. \tag{3.18}$$

It is obvious that for $\gamma = 0$ the model in (3.18) is identical to the simple Cox PH model. This is the null hypothesis H_0 that should be tested versus the alternative $H_A : \gamma \neq 0$, via a score statistic that will eventually follow a chi-square distribution with p degrees of freedom. Consequently, Quantin et al. (1996) offer potential for a global test, which assesses the validity of the PH assumption for two or more covariates simultaneously.

A fascinating fact about the current method is that (3.18) resembles the model in (3.15) in the two-sample case, as γ approaches 0. Indeed, the Maclaurin series of the exponential function e^x is

$$\sum_{j=0}^{\infty} \frac{x^j}{j!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots$$

and thus, as $\gamma \to 0 \Rightarrow \gamma x \to 0$,

$$e^{\gamma x} = 1 + \gamma x + \frac{(\gamma x)^2}{2} + \frac{(\gamma x)^3}{6} + \cdots \approx 1 + \gamma x.$$

As a result, one can consider that $e^{\gamma x} - 1 \approx \gamma x$ and thus, (3.18) can be written as

$$\lambda_1(t) = \lambda_0(t) \exp[\beta + \gamma z(t)], \tag{3.19}$$

where $z(t) = 1 + \ln \Lambda_0(t)$. In contrast to Breslow et al. (1984), Quantin et al. (1996) suggest estimating the baseline cumulative hazard from Breslow's method (see section 2.4.5). Once again, and according to simulations conducted by the authors, the suggested test seems to perform well under crossing hazards.

Bagdonavičius et al. (2004)

The alternative model here has the form

$$\lambda_i(t) = \lambda_0(t) \exp(\beta' x_i) \left[1 + \exp\{(\beta + \gamma)' x_i\} \Lambda_0(t)\right]^{\exp(-\gamma' x_i) - 1}. \tag{3.20}$$

Therefore, a global (score) test is performed based on the model in (3.20) with the null hypothesis being

$$H_0 : \gamma = 0,$$

(Cox PH model)

versus the alternative

$$H_A : \gamma \neq 0.$$

(Cross-effect model/Crossing hazards)

Bagdonavičius & Levulienė (2019)

When the proportionality of some variables is already established and the inclusion of one more, whose proportionality is suspected, is under consideration, a covariate-specific test should be performed. For this purpose, Bagdonavičius & Levulienė (2019) suggested conducting a score test for the null hypothesis

$$H_0 : \gamma = 0$$

versus the alternative

$$H_A : \gamma \neq 0,$$

where γ is a scalar parameter involved in the model

$$\lambda_i(t) = \lambda_0(t) \frac{e^{\beta' x_i + \Lambda_0(t)\exp(\gamma x_{ij})}}{1 + e^{\gamma x_{ij}}[e^{\Lambda_0(t)\exp(\gamma x_{ij})} - 1]}, \tag{3.21}$$

while β is a $p \times 1$ vector of unknown parameters. Note that the resulting statistic follows a chi-square distribution with 1 d.f. (Bagdonavičius & Levulienė, 2019).

Kraus (2007)

Kraus (2007) checks the proportionality of a specified variable x_j using d smooth functions. It is a score test based on the alternative model

$$\lambda_i(t) = \lambda_0(t) \exp[\beta' x_i + \gamma' \psi(t) x_{ij}] \tag{3.22}$$

where β is a $p \times 1$ vector, γ is a $d \times 1$ vector of unknown parameters and $\psi(t) = (\psi_1(t), \ldots, \psi_d(t))$ is the vector of the smoothing functions. Again, $H_0 : \gamma = 0$, $H_A : \gamma \neq 0$, and the final test statistic follows asymptotically a chi-square distribution with d d.f., under the assumption of proportional hazards. According to Kraus (2007), $\psi_k(t), k = 1, 2, \ldots, d$, have the form

$$\psi_k(t) = \varphi_k \left[\frac{\Lambda_0(t)}{\Lambda_0(\tau)} \right]$$

or

$$\psi_k(t) = \varphi_k \left[\frac{F_0(t)}{F_0(\tau)} \right]$$

where Λ_0 and F_0 are the baseline hazard and the baseline survival time distribution, and τ is the total time period of follow-up. The functions $\varphi_k, k = 1, 2, \ldots, d$, should be bounded and linearly independent. Most popular examples are the orthonormal Legendre polynomials on [0,1] and the cosine basis $\varphi_k(u) = \sqrt{2}\cos(\pi k u)$. There are

many other possibilities, such as various spline bases or $\varphi_k(u) = u^k$ (Kraus, 2007; Pena, 1998a, 1998b).

Hafdi (2021)

Finally, another alternative model for the PH assumption, testing on a single covariate while adjusting for others, is the following:

$$\lambda_i(t) = \lambda_0(t)e^{\beta' x_i}\left[1 + e^{\beta_j x_{ij}}t\right]^{\exp(-\gamma x_{ij})-1}. \tag{3.23}$$

Under $H_0 : \gamma = 0$ the score test statistic follows a chi-square distribution with 1 degree of freedom. Hafdi (2021) showed via simulations that the suggested test is more powerful than other similar score tests when the effect of the covariate under the microscope is not linear.

3.2.4 Score process-based tests

Score process-based tests have been extensively studied and compared by Kvaløy & Neef (2004). Some of them are Anderson-Darling's & Cramer-von Mises' test and the Kolmogorov-Smirnov type-based test suggested by Therneau et al. (1990) and Lin et al. (1993). The definition and theoretical justification of these testing procedures are based on an alternative presentation of the data, that of a counting process.

In order to have a better understanding of these approaches, let x_i be the covariate vector for a subject with label $i \in \{1, 2, \ldots, n\}$, whose failure or censoring time is t_i, and define the counting process $N_i(t)$ and the risk indicator $Y_i(t)$ as follows:

$$N_i(t) = I_{\{t_i \le t,\ \delta_i = 1\}},$$

$$Y_i(t) = I_{\{t_i \ge t\}},$$

where δ_i is the failure indicator for the i-th individual. Notice that the counting process for each subject only takes the values 0 and 1, since recurrent event analysis is not under consideration in this thesis.

After having defined $N_i(t)$ and $Y_i(t)$, the p partial likelihood score functions that utilize the information accumulated until time t can be written as

$$U_j(\beta, t) = \sum_{i=1}^{n} \int_0^t \{x_{ij} - \overline{x}_j(\beta, u)\} dN_i(u), \tag{3.24}$$

where

$$\overline{x}_j(\beta, u) = \frac{\sum_{k=1}^{n} Y_k(u) x_{kj} \exp(\beta' x_k)}{\sum_{k=1}^{n} Y_k(u) \exp(\beta' x_k)}.$$

Despite the fact that (3.24) is different from (2.9) at first glance, it holds that $U_j(\beta, \infty) = \partial\ell(\beta)/\partial\beta = U_j(\beta)$ (Therneau et al., 1990; Kvaløy & Neef, 2004; Hafdi, 2021). As for the second derivatives of the partial likelihood, one can define the information matrix for the available information until time t as

$$I(\beta, t) = \sum_{i=1}^{n} \int_0^t V(\beta, u) dN_i(u)$$

where

$$V(\beta, u) = \frac{\sum_{i=1}^{n} Y_i(u) \exp(\beta' x_i)[x_i - \overline{x}(\beta, u)][x_i - \overline{x}(\beta, u)]'}{\sum_{i=1}^{n} Y_i(u) \exp(\beta' x_i)} \tag{3.25}$$

is the weighted covariance matrix of x at time u, and $\overline{x}(\beta, u) = (\overline{x}_1(\beta, u), \ldots, \overline{x}_p(\beta, u))$. Once again, $I(\beta, \infty) = I(\beta)$, i.e., $I(\beta, \infty)$ is equal to the well-known Fisher information matrix for the data at hand. Finally, one last quantity needed for the comprehension of the score process-based tests, is a sequence of p values given by the formula

$$q_j(t) = \frac{I_{jj}(\beta, t)}{I_{jj}(\beta)}$$

where I_{jj} denotes the j-th diagonal element of matrix I. All the aforementioned quantities should be calculated after substituting β with the MPLE $\hat{\beta}$ of the Cox PH model so as to implement any of the tests below.

Kolmogorov-Smirnov type statistic

Therneau, Grambsch & Fleming (1990) and Lin, Wei & Ying (1993) proposed to use this statistic. It tests deviations from proportionality for the j-th variable via the formula

$$KS = \sqrt{\widehat{Var}(\hat{\beta})} \sup_t |U_j(\hat{\beta}, t)|.$$

On a 5% level of significance the null hypothesis is rejected when $KS \geq 1.36$. Kvaløy & Neef (2004), as well as Hafdi (2021), showed via simulations that the Kolmogorov-Smirnov type test is quite conservative, especially in comparison to the other two presented in the current section and some tests from the Grambsch & Therneau family (see section 3.2.5 for more). It also requires orthogonality to provide valid results, meaning that the covariates should be independent. In practice, this is not always the case but small departures from this assumption do not cause great harm.

Cramér-von Mises type statistic

The statistic used here is

$$CV = \widehat{Var}(\hat{\beta}) \int_0^{\infty} U_j(\hat{\beta}, t)^2 d\hat{q}_j(t)$$

On a 5% level of significance the null hypothesis of proportional hazards is reject for the j-th variable when $CV \geq 0.461$. Cramér-von Mises type statistic is considered to have quite good properties against any type of deviation from proportionality, so, it is in a sense, an omnibus test.

Anderson-Darling type statistic

Anderson-Darling score process-based test is believed to be the most powerful amongst other tests in this category, against many types of departure from proportionality. It is in fact a variant of Cramér-von Mises type statistic, given by

$$AD = \widehat{Var}(\hat{\beta}) \int_0^{\infty} \frac{U_j(\hat{\beta}, t)^2}{\hat{q}_j(t)[1 - \hat{q}_j(t)]} d\hat{q}_j(t).$$

Simulation results presented in several papers (Kvaløy & Neef, 2004; Kraus, 2007; Hafdi, 2021) indicated that AD achieves great power when PH assumption does not hold, even under the alternative scenario of non-monotonic HR. Nevertheless, a drawback of this approach is that AD places more weight than CV on regions with possibly few observations. On a 5% level of significance, the null hypothesis is reject when $AD \geq 2.492$.

3.2.5 Grambsch & Therneau's general framework

Undoubtedly, Grambsch and Therneau's (GT) approach for testing the PH assumption is the most popular amongst the aforesaid methods. Major statistical software packages, such as `R` and `Stata`, implement this method. In fact, Grambsch and Therneau's approach can be considered as a family of tests for the assumption of proportionality: a wide range of tests already presented are equivalent to some of its forms under specific conditions. At the same time, related plots can be used to offer an intuitive aspect of the outcome, making the GT family of tests even more appealing (Grambsch & Therneau, 1994; Therneau & Grambsch, 2000).

To gain more insight into the GT family, one must firstly recall the form of the extended Cox model, which generally can be expressed as

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta(t)' x_i\}, \tag{3.26}$$

where $\beta(t) = (\beta_1(t), \beta_2(t), \ldots, \beta_p(t))$ is a $p \times 1$ vector of time-varying coefficients and x_i is the vector of covariates for the i-th individual, as always. Of course, under proportional hazards, $\beta(t)$ must be equal to a vector β consisting of constant, time-invariant values. If this is the case, a plot of $\beta(t)$ versus a function of time or time

itself will reveal a straight line with zero slope, i.e, a horizontal line parallel to x-axis. However, to do the aforesaid plot, one must estimate $\beta(t)$. Grambsch & Therneau (1994) showed that if $\hat{\beta}$ is the coefficient from a typical fit of the Cox PH model, then

$$\beta_j(t_k) \approx \hat{\beta} + E(s^{\star}_{kj}), \tag{3.27}$$

for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, m$. Here, $t_1 < t_2 < \ldots < t_m$ are the ordered failure times and $s^{\star}_{kj}$ are the scaled Schoenfeld residuals.

To fully comprehend the nature of $s^{\star}_{kj}$'s, one must firstly become familiar with the definition of the (unscaled) Schoenfeld residuals. Schoenfeld (1982) introduced these quantities so as to allow the implementation of a GOF test of the Cox PH model in an easily interpretable manner and without the need of complex calculations. The famous Schoenfeld residuals are defined for each variable in the model and for every individual with an observed event during the follow-up period. So, using the previous notation, if m events have been recorded and there are p covariates involved in the assumed model, then the quantities

$$s_{kj} = x_{kj} - \frac{\sum_{\ell \in R_k} x_{\ell j} \exp(\beta' x_\ell)}{\sum_{\ell \in R_k} \exp(\beta' x_\ell)},$$

for $k = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, p$, are known as Schoenfeld residuals. Notice that s_{kj}'s are in fact the elements of the sum that appears in the partial likelihood score equations and in order to be calculated, β should be substituted with the corresponding MPLE $\hat{\beta}$. This is why Schoenfeld initially referred to them as partial residuals. He proposed testing the PH assumption by plotting them against time, for each covariate. Under H_0 it holds that $E[s_{kj}] = 0$ and a plot of $\hat{s}_{kj}$ versus time should be centered around the horizontal line $y = 0$. Under H_A, the alternative of non-PH, one should expect to observe a trend on the values of residuals as time passes. Four years later, a similar, but formal, statistical test was developed by Harrell & Lee (1986), who suggested examining whether the correlation between Schoenfeld residuals and ranked failure time is statistically significant. Under proportionality, it is expected that these quantities are uncorrelated. This is also, a rather popular test and the fact that it is based on the Schoenfeld residuals showcases their importance and usefulness.

Going back to the GT family of tests, it is noticeable that instead of the partial residuals, (3.27) employs a weighted/scaled version of them. The scaled version is given by

$$s^{\star}_{kj} = \hat{V}_k^{-1} s_{kj},$$

where $\hat{V}_k = V(\hat{\beta}, t_k)$ occurs by substituting β with $\hat{\beta}$ and t with t_k in the weighted covariance matrix of x at a specific time point (see (3.25)). Having calculated these residuals, the researcher is ready to perform a test of zero slope on the plot of $\hat{\beta}_j + s^\star_{kj}$ versus a function $g_j(t)$ of time. The visualization alone often provides great understanding of the nature and the extend of non-PH and of course if the test yields in favor of a non-zero slope, it is evidence against PH. Nevertheless, the results should be carefully interpreted as the described test is not omnibus: a specific alternative is being under consideration depending on the choice of $g_j(t)$.

What has been presented as a graphical approach up to now, can be translated into a formal equivalent test, if the elements of the coefficient vector $\beta(t)$ in (3.26) are expressed in the following form:

$$\beta_j(t) = \beta_j + \theta_j(g_j(t) - \overline{g}_j) \tag{3.28}$$

for $j = 1, 2, \ldots, p$. Here $g_j(t)$ is a specified function of time corresponding to the j-th coefficient and $\overline{g}_j$ is equal to $\sum_k g_j(t_k)/m$, i.e., the mean of $g_j(t)$ values over all failure times. The quantities β_j and θ_j are the unknown parameters of the assumed model, and naturally the null hypothesis of PH can be written as

$$H_0 : \theta_j = 0, \forall j \in \{1, 2, \ldots, p\},$$

while the alternative is

$$H_A : \theta_j \neq 0, \text{ for at least one } j \in \{1, 2, \cdots, p\}.$$

Of course under H_0, β_j is estimated as the MPLE of the Cox PH model. Consequently, combining (3.27) and (3.28), an interesting approximate relationship for the mean of the scaled Schoenfeld residuals occurs:

$$E[s^\star_k] \approx G_k\theta \tag{3.29}$$

where G_k is a $p \times p$ diagonal matrix whose (j, j) element is $g_j(t_k) - \overline{g}_j$, and $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$. Also, it holds that their variance is

$$Var[s^\star_k] \approx \hat{V}_k^{-1} - \left[\sum_{l=1}^{m} \hat{V}_l\right]^{-1} \tag{3.30}$$

for $k = 1, 2, \ldots, m$. Therefore, under H_0, vector θ can be estimated via (3.29) and (3.30), implementing the multivariate generalized least squares (GLS) technique. At last,

$$\hat{\theta} = Q^{-1} \sum_{k=1}^{m} G_k \hat{s}_k$$

where

$$Q = \sum_{k=1}^{m} G_k \hat{V}_k G_k - \left[\sum_{k=1}^{m} G_k \hat{V}_k\right] \left[\sum_{k=1}^{m} \hat{V}_k\right]^{-1} \left[\sum_{k=1}^{m} G_k \hat{V}_k\right]'$$

and the test statistic for the null hypothesis is

$$T(G) = \hat{\theta}' Q \hat{\theta} = \left[\sum_{k=1}^{m} G_k \hat{s}_k\right]' Q^{-1} \left[\sum_{k=1}^{m} G_k \hat{s}_k\right]. \tag{3.31}$$

$T(G)$ has an asymptotic chi-square distribution with p d.f. when the PH assumption holds and it can be used for a global test which also has a graphical interpretation for each covariate.

Notice that the value of the statistic in (3.31) depends on the choice of the time function. Undoubtedly, different choices of G result in different tests for model misspecification. What is intriguing, though, is that depending on the form of $g(t)$, equivalent tests to the ones presented in previous sections occur[6]. More specifically, according to Grambsch & Therneau (1994), if

1. $g(t)$ is a specified function of time, such as t or $\ln t$, then $T(G)$ is a score test for the addition of the time-varying covariate $g(t)\dot{x}$ to the model. This test was initially suggested by Cox (see section 3.2.3). It also seems to have a connection with Gill & Schumacher's proposal (see section 3.2.2), since according to Chappell (1992) the latter approach is a variant of Cox's which, however, imposes unnecessary limitations[7].

2. $g(t)$ is a piecewise constant function on non-overlapping time intervals, $T(G)$ is the score test suggested by O'Quigley & Pessione (see section 3.2.1). Since the latter generalizes and extends the tests proposed by Schoenfeld (1980) and Moreau et al. (1985), it can be assumed that all three of them belong to the GT family. Time intervals should be determined before the analysis in order to obtain valid and unbiased results. Some suggestions about the choice of time partition are given in previous sections and are mostly based on the papers published by Andersen (1982) and Schoenfeld (1980).

3. $g(t)$ is the number of events until time point t, then $T(G)$ is the covariance between the scaled Schoenfeld residuals and the ranked failure times. This is

[6] In everything mentioned afterwards, G is a diagonal matrix and its non-zero elements are equal to the same value $g(t)$.

[7] Gill & Schumacher's method is appropriate only for the two-sample case, whereas Cox's proposal is easily extendable to the multi-sample case.

almost the description of Harrell & Lee's test[8]. It is also equivalent to the test of Breslow et al. (see section 3.2.3) since it uses rank scores in (3.26).

4. $g(t)$ is a weighting function from the FH family (or any other family of weights), then $T(G)$ mimics, in a sense, the test suggested by Lin (see section 3.2.2). In fact, if the weighted estimate $\hat{\beta}_w$ provided by Lin's approach occurred from a one-step Newton-Raphson algorithm starting from $\hat{\beta}$, then these two tests would be identical.

5. $g(t)$ is equal to the lagged residuals, i.e., $g_j(t_1) = 0$ and $g_j(t_{k+1}) = a_j^2 \hat{s}_{kj}$ for $j = 1, 2, \ldots, p$, then $T(G)$ gives a test suggested by Nagelkerke, Oosting & Hart (1984). Essentially, it is a proportionality test using the serial correlation of the Schoenfeld residuals for a univariate predictor, or the correlation of a weighted sum $a'\hat{s}_k$ for the multivariate case. Usually $a = \hat{\beta}$. No further information will be given about this approach, since it has been proved to lack power and perform poorly in comparison to other tests presented in the current thesis (Quantin et al., 1996).

It is evident that a wide variety of tests proposed in the literature can be expressed as a $T(G)$ test from the GT family, and that explains its popularity.

Summarizing this section, Grambsch & Therneau (1994) derived a test for proportionality which can be roughly thought of as a test of zero slope in a regression line fit to a plot of the scaled Schoenfeld residuals against a time function $g_j(t)$. Both a global test of proportionality and separate tests for each covariate are provided (if the functions $g_j(t)$ are selected accordingly). Different choices of $g_j(t)$ correspond to different tests, i.e., tests with different alternatives, and several earlier proposed tests of proportionality are special cases of this family, corresponding to particular choices of the time function. A limitation with the GT family of tests is that only one specific alternative to proportional intensity, namely time-dependent coefficients, is checked. Other kinds of deviations can possibly be wrongly interpreted or not detected at all. Another limitation is the need to choose specific time functions. This may lead to low power against deviations of a kind not described by this function, for instance, a non-monotonic deviation when a monotonic $g_j(t)$ function has been chosen.

[8]Despite the fact that the original test includes the unscaled Schoenfeld residuals, numerous papers support that the results should be similar whether someone uses the weighted or the unscaled version.

To complete the section of formal statistical tests for the proportionality assumption, Table 3.1 provides an enlightening categorization of them into three groups: global, univariate, and two-sample tests. Of course, global tests can be modified to test the proportionality for a single covariate and thus, become univariate, and at the same time, all univariate tests can be implemented for a dichotomous covariate becoming two-sample tests. Nevertheless, Table 3.1 helps the reader understand better the capabilities of each test described in this section.

Global	Univariate	Two-sample
Schoenfeld (1980)	Andersen (1982)	Gill & Schumacher (1987)
Moreau et al. (1985)	Bagdovaničius & Levulienė (2019)	Sengupta et al. (1998)
O'Quigley & Pessione (1989)	Hafdi (2021)	Breslow et al. (1984)
Lin (1991)	Kraus (2007)	
Cox (1972)	Cramér-von Mises (2004)	
Bagdonavičius et al. (2004)	Anderson-Darling (2004)	
Nagelkerke et al. (1984)	Kolmogorov-Smirnov (1990)	

Table 3.1: Classification of tests for proportional hazards.

3.3 Graphical tests

An extremely large number of graphical approaches for testing the PH assumption has been developed since the introduction of the Cox model in 1972. Graphical tests are really helpful when the number of covariates is small and quite informative in the presence of qualitative variables with few categories. However, today is the era of Big Data and thus, the occasions on which such simple problems occur are rare. That is the main reason why graphical tests are not frequently used, along with the fact that their interpretation is rather subjective. The examination of graphs is not an easy task and requires knowledge and experience. Unfortunately, even when a statistician acquires these skills it is possible to misinterpret them.

To achieve consensus among results from analyses conducted by different statisticians, it is suggested that findings are based on formal tests rather than arbitrary interpretations of graphs. Consequently, many papers report p-values and avoid displaying figures to justify some of the results or the choice of methodology. In any case, it is crucial to remember that this general agreement does not downplay the importance of graphical tests. On many occasions, graphical tests complement the formal ones, verify their results, and provide some sort of guidelines, facilitating the subsequent steps of analysis. Therefore, even if their role is mostly complementary, statisticians must acclimatize to the most commonly used in practice. Some of them

are presented in the following sections.

3.3.1 Based on residuals

Residual plots have been extremely helpful throughout the years, not only in Survival Analysis but also in the general field of Statistics. Think, for instance, all the tests performed to evaluate the fit of the simplest and most popular statistical model: the simple linear regression. Residual plots give great insight into the nature of the data and the relationships between covariates and response. It would be out of the ordinary not to use an analogous approach for the evaluation of the fit of the Cox model and consequently, for the examination of the validity of the PH assumption. Already, in section 3.2.5, during the presentation of the GT family of tests, a graphical equivalent approach based on the Schoenfeld residuals has been described. However, since there are also other methods based on residuals, this will be presented briefly here, and it will be followed by two new graphical tests.

Schoenfeld residuals versus time

As mentioned before, after calculating the scaled or unscaled Schoenfeld residuals for each variable x_j, a statistician can create a plot of $\hat{s}_{kj}$ (or $\hat{s}^{\star}_{kj}$) versus a function of time or time itself. Under the PH assumption, there should not be a trend, meaning that if a line is fitted to the graph it should roughly have zero slope. This means, that under proportionality, the residuals should form a random "cloud" around the time axis (x-axis). On the other hand, under the alternative of non-PH, the choice of time function will determine the shape of the residuals in the plot. However, an incorrectly specified function $g(t)$ may result in a plot where no trend appears. Usual choices for g are the (ranked) time itself, the natural logarithm of time and the KM estimator of the survival function based on the whole dataset.

Cumulative sums of Schoenfeld residuals versus time

Another graphical approach is to use the cumulative sums of Schoenfeld residuals against scaled time to (0,1), for each covariate. Under PH, each curve should be a Brownian bridge, i.e., a random walk starting and ending at 0.

Kay's residuals

Last but not least, another type of residuals can be used for the assessment of

the PH assumption. Kay (1977) defined the residuals

$$\varepsilon_i = \int_0^{t_i} \lambda_0(u) \exp(\beta' x_i) du,$$

where t_i are the failure or censoring time of the i-th individual. For the calculation of ε_i, β and λ_0 are substituted with their estimates. Under H_0, these quantities should exhibit approximately the properties of a random sample of size n from a unit exponential distribution. Therefore, the model fit can be checked via a plot of the estimated cumulative hazard of the observed residuals ε_i. If PH holds, the plot should show a straight line passing through the origin with slope unity.

3.3.2 Based on cumulative hazard plots

Apart from the residual plots, cumulative hazard graphs are also particularly helpful, not only for testing the PH assumption but also for selecting a parametric model which may have the potential of describing the data in a better way than a semi-parametric model. Notwithstanding this advantage, cumulative hazard plots are mostly used for the two-sample case, since a generalization for more covariates (or levels of a qualitative variable) is somewhat complex to be visualized. Therefore, if there are two groups of interest in the data with corresponding cumulative hazard functions Λ_1 and Λ_2, usually one of the following five graphs is created[9]:

1. $\ln \Lambda_1(t)$ and $\ln \Lambda_2(t)$ versus time: Under the PH assumption it holds that

$$\frac{\lambda_2(t)}{\lambda_1(t)} = \theta \Rightarrow \frac{\Lambda_2(t)}{\Lambda_1(t)} = \theta$$

for some positive constant θ. Therefore, taking the natural logarithms in the two parts of the last equation, it occurs that $\ln \Lambda_2(t) - \ln \Lambda_1(t) = \ln \theta$. Since the difference of the logarithms of the cumulative hazards is constant, the plot of $\ln \Lambda_1(t)$ and $\ln \Lambda_2(t)$ versus time should display two parallel lines under H_0. If that is not the case, e.g. if the curves cross each other, then proportionality is rejected.

An interesting fact here is that this plot is identical to a rather famous one, called log-minus-log plot which is based on the survival curves of the two groups. According to this approach, the curves $\ln[-\ln S_1(t)]$ and $\ln[-\ln S_2(t)]$ should be plotted against time. However, from (2.2) it is obvious that the two curves are exactly the same with $\ln \Lambda_1(t)$ and $\ln \Lambda_2(t)$, respectively.

[9]For all the plots, the Nelson-Aalen estimator of the cumulative hazards is preferred.

2. $\ln \Lambda_2(t) - \ln \Lambda_1(t)$ versus time: according to what was mentioned above, this difference should be constant as time passes when PH holds. Consequently, under proportionality, a straight horizontal line is expected. Naturally, this plot is called *log cumulative hazard difference plot* or *LCHD plot* (Dabrowska et al., 1992).

3. $\frac{\Lambda_2(t)-\Lambda_1(t)}{\Lambda_1(t)}$ versus time: This is the *relative cumulative hazard difference plot* or *RCHD plot*. Under proportionality, it holds that

$$\frac{\Lambda_2(t) - \Lambda_1(t)}{\Lambda_1(t)} = \frac{\Lambda_2(t)}{\Lambda_1(t)} - 1 = \theta - 1,$$

for a positive constant θ. Consequently, if the PH assumption is valid, a horizontal line is expected to appear in the corresponding plot (Dabrowska et al., 1989).

4. $\Lambda_2(t)$ versus $\Lambda_1(t)$: Under the assumption of proportional hazards it holds that

$$\frac{\Lambda_2(t)}{\Lambda_1(t)} = \theta \Rightarrow \Lambda_2(t) = \theta \Lambda_1(t)$$

for a positive constant θ and $\forall t \in [0, \tau]$, where τ is the maximum observed failure or censoring time. Thus, when proportionality holds, someone would anticipate to see a straight line through origin with slope θ.

5. $\Lambda_2[\Lambda_1^{-1}(u)]$ versus $u, 0 < u < \Lambda_1(\tau)$: this method was initially proposed by Lee & Perie (1981) and the function used is called the *trend function*. This plot is in fact identical to the previous one but for different values of t. This means that under the PH assumption, a straight line through the origin with slope equal to the real HR should appear in the plot. Under the alternative of monotonic HR, one should expect to see a convex or concave curve, when the HR is increasing or decreasing over time, respectively, due to the fact that the first derivative of the trend function is equal to the HR. In the literature, this plot is referred to as *relative trend function plot* or *RTF plot*.

Despite their simplicity and usefulness, most of these plots have been characterized as unstable especially for small samples, since they tend to have wild fluctuations at the beginning of time or may lack precision for large values of t (Sengupta et al., 1998; Sahoo & Sengupta, 2016). To overcome this problem, it was proposed to use their weighted counterparts, replacing $\Lambda_j(t), j = 1, 2$, with $T_j(t)$, where

$$T_j(t) = \int_0^t K(s)\Lambda_j(s)ds$$

and $K(s)$ is a weighting function from the FH family. Sengupta et al. (1998) have stressed that the relative plots are smoother and more stable even for small sample sizes. At the same time, Sahoo & Sengupta (2016) pointed out that a monotone decreasing function can bring more stability.

Finally, some attempts have been made throughout the years, to combine the interpretability of the graphical tests with the formality of the analytical ones. With this aim, some approaches relied on confidence bands. For instance, in the literature, it is sometimes suggested to create an RTF or LCHD or RCHD plot (or their weighted counterparts) and check if the band contains a straight line through origin for the first plot and if an horizontal line fits in the asymptotic confidence bands for the other two graphs (Dabrowska et al., 1989, 1992). Nevertheless, these tests have low power and therefore, other methods have been developed in a similar spirit. The most famous amongst them is a test proposed by Sahoo & Sengupta (2016) and it is based on acceptance bands. Despite the fact that it is a rather complex procedure, it combines the good power of analytical tests with a graphical visualization. It also captures various types of departure from proportionality.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 4

Simulation study: Tests for proportional hazards

4.1 Previous simulation studies

In Chapter 3 a great variety of tests for proportional hazards has been presented. A considerable number of papers comparing these tests has been published, yet there is not a generally proposed approach, or a test that is robust against various deviations from proportionality. The main reason for this issue is that it is impossible to check all potential scenarios of non-PH, while at the same time, most of the tests are disregarded due to nonexistence of relevant statistical software packages. In `R`, `Stata`, and `SAS`, the user can apply either the GT family of tests or Cox's suggestion of creating a time-varying variable and checking its significance in the model. Despite the fact that the first method is quite flexible allowing for a simultaneous test of multiple covariates, and the second produces an extended model for the data at hand, the research on this particular problem is so rich that it would be naive to conclude that GT family and Cox's test are the optimal choices against every pattern of non-PH.

Some attempts for comparison of different testing procedures have been made already: Song & Lee (2000) have shown that Gill and Schumacher's (1987) test performs well when the HR is monotone, whereas Schoenfeld's (1980) and Andersen's (1982) interval dependent tests seem to be more appropriate under non-monotonic and irregular patterns of non-PH. Quantin et al. (1996) compared a great variety of tests for the two-sample case, using increasing and decreasing hazard ratio functions and simulating data from the Weibull distribution. Again, Gill & Schumacher's test achieved great power along with the proposed test, while Breslow, Elder & Berger's (1984) proposal using the cumulative hazards score was very close in terms

of performance (Quantin et al., 1996). On the other hand, Sahoo & Sengupta (2016) examined the performance of Gill & Schumacher's test versus Wei's (1984) proposal, which is essentially a score process-based test similar in spirit to Lin's (1991) in section 3.2.2, and came to the conclusion that the first has poor power in comparison to the second when the HR is a bathtub-shaped function, despite their equally good performance for large samples under monotonous deviations from proportionality. Lastly, the performance of various score process based tests, such as Anderson-Darling, Cramér-von Mises and Kolmogorov-Smirnov was investigated along with other tests either from the GT family or alternative model specifications, from Hafdi (2021) and Kvaløy & Neef (2004). The Kolmogorov-Smirnov type test seemed to be the most conservative, while the Anderson-Darling type test had somewhat inflated type I error when the variables under consideration were highly correlated. However, the latter achieved great power when the HR was a non-monotonous function of time, and under monotonous patterns as well. At the same time the performance of GT tests using the rank of the failure times or the natural logarithm of time as $g(t)$, has been questioned, at least in comparison to the score process-based tests.

Unfortunately, it has been acknowledged that there is not a general consensus about which test is better under different types of non-proportionality, partly because the possibilities are endless. A single test cannot be powerful against all situations, and so, practical consideration should be taken into account when deciding which procedure to use. Since a-priori knowledge of the probable type of non-PH pattern is rare, some claim that applying several tests simultaneously will give some protection against misspecified alternatives (Song & Lee, 2000).

4.2 Data simulation: Special scenarios

Numerous alternatives for the nature of the data and the non-proportionality patterns can be considered. In the current thesis, the simulation study will focus on the two-sample problem, which is of great concern when it comes to the analysis of survival data from clinical trials. The comparison of two treatments is oftentimes the main subject of such an analysis and questions about the superiority of one over the other should be addressed based on valid results. If the Cox PH model is fitted to the data in order to obtain a summary measure for the relative risk, or the log-rank test is implemented, the PH assumption must be tested, otherwise the conclusions might be misleading.

For the simulation of the data, the piecewise exponential distribution is used,

since, according to Lin et al. (2020), this particular distribution tends to mimic the behavior of the data collected in real-life applications. The null hypothesis of proportionality is examined along with the four basic scenarios of non-PH presented at the beginning of Chapter 3: early/diminishing effect, late/delayed effect, crossing hazards and long-term survivors. In all these cases, the hazard function of the control group, i.e., the baseline function $\lambda_0(t)$, is constant and equal to 1. This means that the distribution of the survival time for the placebo group is exponential with rate equal to 1. As for the hazard function $\lambda_1(t)$ of the intervention group, it changes according to each scenario. More specifically,

- For the null hypothesis of proportional hazards, three cases are investigated corresponding to HRs equal to 0.65, 0.80 and 0.90, i.e., $\lambda_1(t) = 0.65$ or $\lambda_1(t) = 0.80$ or $\lambda_1(t) = 0.90$ (see Figure C.1 in Appendix C).

- For the diminishing effect, three different scenarios are considered:

$$\lambda_1(t) = \begin{cases} 0.65, & \text{if } t \leq t_{CP} \\ 0.99, & \text{if } t > t_{CP} \end{cases}$$

$$\lambda_1(t) = \begin{cases} 0.80, & \text{if } t \leq t_{CP} \\ 0.99, & \text{if } t > t_{CP} \end{cases}$$

$$\lambda_1(t) = \begin{cases} 0.90, & \text{if } t \leq t_{CP} \\ 0.99, & \text{if } t > t_{CP} \end{cases}$$

  where t_{CP} is chosen as the time point at which 30%, 50% and 70% of events are expected to happen in the treatment group (3 cases per hazard function, and thus 9 in total; see Figure C.2 in Appendix C).

- For the late effect, data with the following hazard functions are simulated:

$$\lambda_1(t) = \begin{cases} 0.99, & \text{if } t \leq t_{CP} \\ 0.65, & \text{if } t > t_{CP} \end{cases}$$

$$\lambda_1(t) = \begin{cases} 0.99, & \text{if } t \leq t_{CP} \\ 0.80, & \text{if } t > t_{CP} \end{cases}$$

$$\lambda_1(t) = \begin{cases} 0.99, & \text{if } t \leq t_{CP} \\ 0.90, & \text{if } t > t_{CP} \end{cases}$$

and again, t_{CP} is chosen as the time point at which 30%, 50% and 70% of events are expected to happen in the treatment group (9 scenarios in total; see Figure C.3 in Appendix C).

- For the crossing hazards pattern, four different scenarios are being investigated:

$$\lambda_1(t) = \begin{cases} 0.65, & \text{if } t \leq t_{CP} \\ 1.10, & \text{if } t > t_{CP} \end{cases}$$

$$\lambda_1(t) = \begin{cases} 1.10, & \text{if } t \leq t_{CP} \\ 0.65, & \text{if } t > t_{CP} \end{cases}$$

$$\lambda_1(t) = \begin{cases} 0.80, & \text{if } t \leq t_{CP} \\ 1.20, & \text{if } t > t_{CP} \end{cases}$$

$$\lambda_1(t) = \begin{cases} 1.20, & \text{if } t \leq t_{CP} \\ 0.80, & \text{if } t > t_{CP} \end{cases}$$

  where $t_{CP} = 0.7$. This specific time point is chosen because by $t = 0.7$ almost half of the events in the whole dataset are anticipated. For each scenario of the above, the study is assumed to end at $\tau_{\text{end}}^{(1)} = 2$ and $\tau_{\text{end}}^{(2)} = 4$ and thus, an approximate additional 30% and 45% of events are expected, respectively (8 scenarios in total; see Figure C.4 in Appendix C).

- Finally, for the case of long term survivors, only two scenarios are considered:

$$\lambda_1(t) = \begin{cases} 0.65, & \text{if } t \leq t_{CP} \\ 0.65^2, & \text{if } t > t_{CP} \end{cases}$$

$$\lambda_1(t) = \begin{cases} 0.80, & \text{if } t \leq t_{CP} \\ 0.80^2, & \text{if } t > t_{CP} \end{cases}$$

  and again $t_{CP} = 0.7$. In each case, the study is assumed to end at two different time points $\tau_{\text{end}}^{(1)} = 2$ and $\tau_{\text{end}}^{(2)} = 4$, resulting in 4 approaches in total (see Figure C.5 in Appendix C).

Apart from the administrative censoring in the crossing hazards and long-term survivors scenarios, random censoring is also assumed. The censoring time for each subject is independent of the group it belongs to and follows an exponential distribution with a rate which will result in a 5% of censored observations in the absence of fixed censoring. Moreover, the sample size is set equal to $n = 200$ and $n = 1000$ with half of the patients in group 1 and the rest in group 2, leading to 66 types of simulated data! The number of repetitions used is equal to 1000 for each scenario.

4.3 Results

Eighteen of the tests presented in Chapter 3 are compared in this section:

1. Cox's (1972) test, adding the interaction of the treatment indicator with time in the PH model,
2. Cox's (1972) test, adding the interaction of the treatment indicator with the natural logarithm of time in the PH model,
3. Cox's (1972) test, adding the interaction of the treatment indicator with a step function of time in the PH model (the function is equal to zero before a certain time point and equal to 1 afterwards),
4. Grambsch & Therneau's (1994) test, using $g(t) = t$,
5. Grambsch & Therneau's (1994) test, using $g(t) = \ln t$,
6. Grambsch & Therneau's (1994) test, using as $g(t)$ the step function described in the third test,
7. Grambsch & Therneau's (1994) test, using as $g(t)$ the ranks of the failure times,
8. Grambsch & Therneau's (1994) test, using $g(t) = \hat{S}(t)$, i.e., the KM estimate,
9. Gill & Schumacher's (1987) test with weights corresponding to the log-rank and Peto-Prentice statistics,
10. Lin's (1991) test using the weights proposed by Schemper et al. (2009),
11. Lin's (1991) test using the weights proposed by Xu & O'Quigley (2000),
12. Schoenfeld's (1980) interval-dependent test with one change point,
13. Moreau, O'Quigley & Mesbah's (1985) proposal for a conservative counterpart of the previous test,
14. Andersen's (1982) test with two intervals,
15. O'Quigley & Pessione's (1989) test with two intervals and $\Psi_1 = 1, \Psi_2 = -1$,
16. Breslow, Elder & Berger's (1984) approach with rank scores,

17. Breslow, Elder & Berger's (1984) approach with cumulative hazards scores, and lastly,

18. an approximation of the test suggested by Quantin et al. (1996) using the previous test with scores equal to $1 + \ln \Lambda_0(t_i)$ for each event time $t_i, i = 1, 2, \ldots, m$.

The results per each special scenario are discussed below. Additional tables and figures for a better understanding and justification of the findings can be found in the Appendix Section A.

Proportional Hazards

Table 4.1 displays the results for the 18 aforementioned tests under three possible values for the HR of the intervention versus the control group when the sample size is either small ($n = 200$) or large ($n = 1000$). According to this, when the proportionality assumption is valid, the empirical significance level appears to be close to the nominal level 5% in most of the situations studied for every test.

	Hazard Ratio					
	0.65		0.8		0.9	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	8.6	6.3	7.3	6.3	6.6	5.4
2	4.9	5.0	4.9	4.8	5.4	5.4
3	5.9	4.8	5.7	4.5	6.3	4.5
4	6.4	5.6	5.6	5.5	5.3	5.1
5	4.7	4.9	4.9	5.0	5.0	5.3
6	5.9	4.7	5.8	4.3	6.0	4.8
7	5.2	5.1	5.4	4.8	5.5	4.8
8	5.1	5.1	5.4	4.7	5.4	4.8
9	4.8	4.7	4.8	3.8	4.9	4.0
10	5.3	5.1	5.2	5.1	5.4	5.2
11	2.8	4.5	3.7	5.0	4.1	5.3
12	5.6	4.5	6.0	4.5	6.3	4.4
13	5.6	4.3	5.5	4.4	6.0	4.3
14	5.1	4.5	6.2	4.3	5.7	4.9
15	5.6	4.5	6.0	4.5	6.3	4.4
16	5.2	5.2	5.4	4.7	5.6	4.8
17	6.3	5.3	5.8	5.3	5.6	5.3
18	4.6	4.4	4.8	4.9	4.9	5.1

Table 4.1: Type I error (size in %) of 18 tests for proportional hazards in the two-sample case, using three constant HR functions and two different sample sizes n.

Test 1 has somewhat inflated type I error, especially for small samples. This is a Wald test for the significance of the interaction of time with the treatment indicator. Even though Test 4 is an equivalent procedure from the GT family (score test for the same interaction term), it is evident that the latter is more conservative and thus, more valid under proportional hazards, but the empirical significance level is still inflated. As for the tests based on weighting functions, Gill & Schumacher's suggestion (Test 9) is quite conservative, especially for large samples, and the same holds for Lin's test (Test 11) with weights from Xu & O'Quigley (2000), but for small samples. Finally, another interesting finding which confirms a statement about the relationship of Schoenfeld's and Moreau, O'Quigley & Mesbah's proposal, is that the latter (Test 13) is indeed more conservative than the first (Test 12). Figure 4.1 complements these results and comments, offering a graphical justification of what was reported up to now.

Early/Diminishing Effect

Table 4.2 and Figure 4.2 offer great insight into the performance of the 18 aforementioned tests for different sample sizes and change points, when an early effect with initial HR = 0.65 is observed. More specifically, if the HR changes when 30% of events have been occurred, rank and KM tests from the GT family seem to perform better, along with Breslow's test using the rank scores. Their performance remains comparable with other tests if the change happens at 50% of events. On the other hand, the GT test with $g(t) = t$ and the equivalent Cox test, along with Breslow's cumulative hazards score test and Lin's with Xu & O'Quigley's weights, lack power in both situations regardless the sample size. Interestingly, they exhibit the highest power when the change takes place after the occurrence of an approximate 70% of the events in the treatment group. In general, the interval-dependent tests (Tests 3, 6, and 12 to 15) display mediocre performance which reaches its crescendo when $x = 50\%$. This is rather reasonable and anticipated since these tests compare the behavior of the data before a certain time point with their behavior after that. This time point is chosen so as to split the times axis into two intervals containing similar numbers of events. Nevertheless, even in this case, interval-dependent tests do not perform better than GT rank, GT KM and Breslow's rank score test. A general comment on Table 4.2 is that almost all tests perform better when the change in the HR happens at the beginning or in the middle of the follow-up.

Tables A.1 & A.2, and Figures A.1 & A.2 in the Appendix A show the relative results from the other two scenarios of early treatment effect (initial HR = 0.8 and

Figure 4.1: Type I error (size) of 18 tests for proportional hazards, for each sample size and HR. The dashed line corresponds to type I error equal to 5%.

0.9). The findings are quite similar, but the power of all tests is severely diminished since the deviation from proportionality is not so evident.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	17.2	50.4	21.2	69.3	15.5	52.3
2	17.7	66.6	14.5	57.3	8.1	27.0
3	18.5	66.9	20.0	70.4	8.9	23.3
4	15.4	47.7	19.6	65.3	14.6	49.7
5	18.7	67.2	16.1	58.7	9.1	28.5
6	17.9	67.0	19.8	70.8	9.3	23.3
7	20.7	73.4	21.1	75.0	11.5	40.7
8	20.3	73.1	21.3	75.1	11.6	41.5
9	19.3	71.2	20.0	72.9	11.0	39.4
10	19.7	71.5	20.8	73.1	11.5	40.9
11	13.5	45.2	17.1	61.4	10.9	48.8
12	17.9	67.0	20.3	70.9	8.9	24.2
13	17.1	66.7	19.7	70.5	8.6	23.9
14	17.7	66.0	20.3	72.0	8.8	28.8
15	17.9	67.0	20.3	70.9	8.9	24.2
16	20.6	73.4	21.2	74.9	11.6	40.5
17	15.0	45.6	18.7	62.5	14.8	48.4
18	18.6	67.8	16.0	61.0	9.4	30.0

Table 4.2: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the early effect case with initial HR = 0.65 and subsequent HR ≈ 1 after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

Late/Delayed Effect

Again, in this case, only the results for the first scenario of final HR = 0.65 are presented in the main body of this thesis, but relative tables and graphs for the other two scenarios can be found on the Appendix Section A (Tables A.3 & A.4, Figures A.3 & A.4). To begin with, when the effect changes at the beginning of the follow-up, GT test with $g(t) = \ln t$ and the equivalent Cox, along with GT rank, GT KM, Gill & Schumacher's and Lin's test with weights proposed by Schemper et al. (2009) seem to exhibit the highest power. When the change happens after the occurrence of 50% of events in the treatment group, interval-dependent tests (Tests 3, 6 and 12 to 15) reach their peak in terms of power, outperforming the aforementioned group. Nevertheless, the power of GT rank, GT KM, Gill & Schumacher's and Lin's test is rather close to the power of the interval dependent tests. An interesting observation is that tests such as GT with $g(t) = t$, the equivalent test by Cox, Breslow's test with cumulative hazard scores and Lin's test using the weights introduced by Xu & O'Quigley (2000) display the worst performance except from the case where the change in HR takes place at the end of the study. In fact, the latter test is not at all reliable when the sample size is small or the effect increases at the beginning of the study. Finally, even in the last scenario (CP = 70%) Tests 7 to 10 achieve power levels close to the best ones (Table 4.3, Figure 4.3). The same holds for the cases when final HR = 0.8 or 0.9 but the power of all tests is severely decreased (approximately 10% for $n = 200$ and mainly 25-35% for $n = 1000$ when HR = 0.8, and 6-7% for $n = 200$ and at most 12.2% for $n = 1000$ when HR = 0.9).

Looking closely, one can notice that the results for the early and late effect scenarios are similar. Consequently, it is safe to say that good options, if such non-PH patterns are excepted, are Tests 7, 8, 9 and 10. They may not outperform all the others under all alternatives, but even when they don't, their performance is comparable to the best choice.

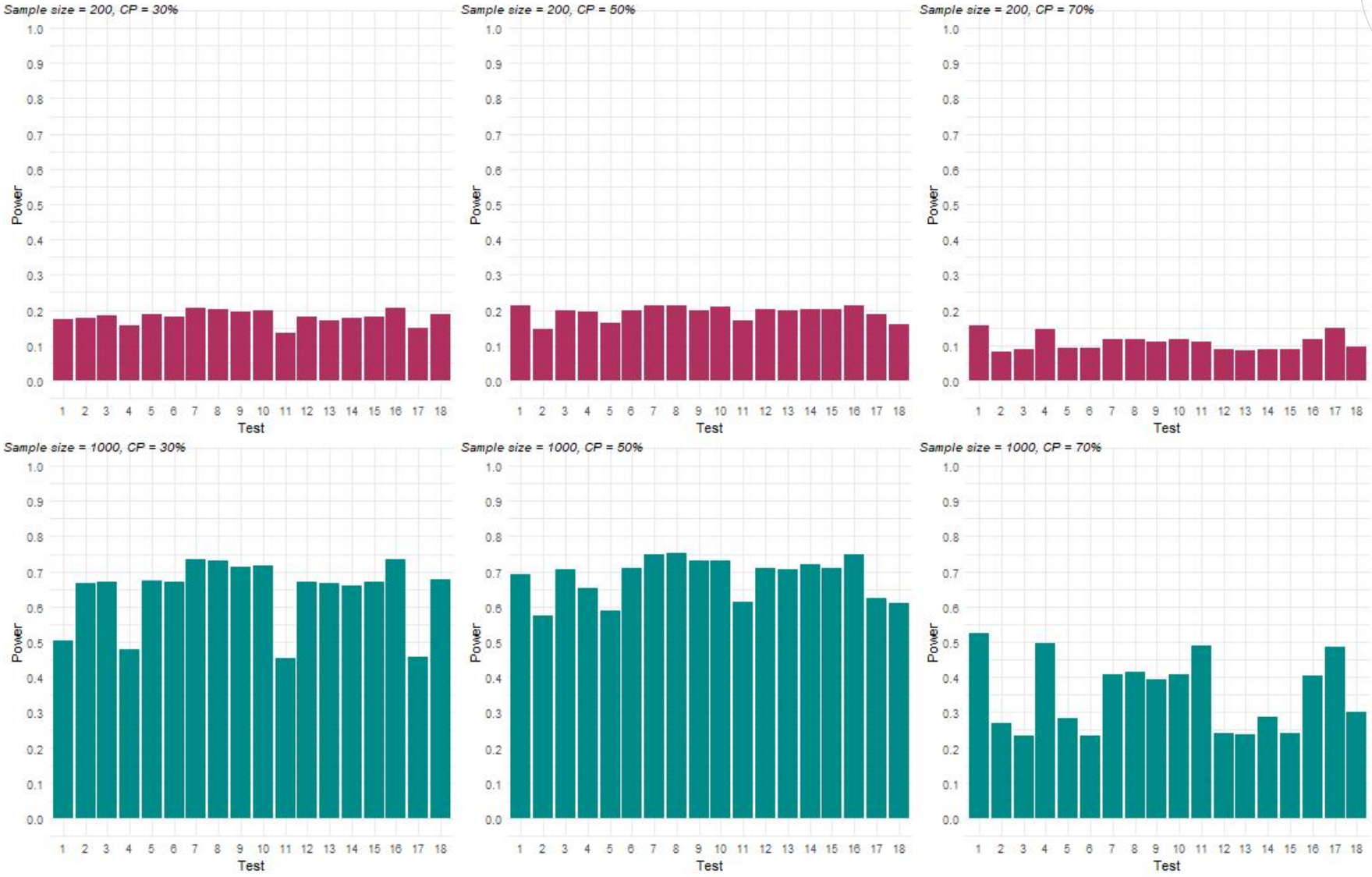

Figure 4.2: Power of 18 tests for proportional hazards, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when an early effect with initial HR = 0.65 is observed.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	20.5	45.0	27.3	62.4	25.3	63.7
2	20.1	67.8	20.0	69.5	13.5	46.6
3	16.4	51.9	27.6	85.6	13.3	42.6
4	18.5	46.9	25.9	66.4	23.5	67.2
5	20.4	68.2	20.1	69.3	13.4	46.3
6	16.2	52.0	27.3	85.7	12.7	42.2
7	22.2	68.4	26.0	80.2	17.6	59.8
8	22.0	68.1	26.6	80.0	17.7	60.7
9	21.6	66.7	25.3	78.6	16.9	58.6
10	22.3	66.8	26.9	78.7	18.3	61.0
11	8.3	42.1	13.2	62.0	14.3	63.6
12	16.0	51.5	27.1	85.5	12.8	42.3
13	14.9	51.0	26.0	85.4	12.1	41.9
14	16.2	52.0	27.3	87.0	15.4	49.6
15	16.0	51.5	27.1	85.5	12.8	42.3
16	22.2	68.6	26.3	79.9	17.6	59.6
17	19.8	49.7	26.3	69.6	23.4	68.8
18	20.0	68.4	20.0	68.3	13.8	44.3

Table 4.3: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the late effect case with initial HR ≈ 1 and subsequent HR $= 0.65$ after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

Figure 4.3: Power of 18 tests for proportional hazards, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when a late effect with final HR = 0.65 is observed.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	28.4	88.5	25.7	83.1
2	20.5	76.0	24.0	83.9
3	35.6	95.2	35.6	95.2
4	28.6	88.5	24.9	82.8
5	23.2	77.7	26.0	84.6
6	35.2	95.1	36.1	95.8
7	31.1	90.0	32.0	91.9
8	31.1	90.2	31.5	91.8
9	11.3	71.3	29.8	89.4
10	29.7	89.5	31.1	90.8
11	26.9	89.2	21.4	80.7
12	35.4	95.2	35.3	95.2
13	35.4	95.2	34.8	95.2
14	34.9	95.2	34.5	95.4
15	35.4	95.2	35.3	95.2
16	31.1	90.2	31.9	91.8
17	28.3	88.1	23.5	80.3
18	24.2	80.4	26.3	85.9

Table 4.4: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the scenario of crossing hazards with initial HR = 0.65 and subsequent HR = 1.10, for different sample sizes n and cut points 2 and 4.

Crossing Hazards

Table 4.4 presents the related outcome for the case where the initial HR is equal to 0.65 and then an inversion of effect is observed, with a new HR = 1.10. It is evident that Gill & Schumacher's and Lin's test (with Xu & O'Quigley's weights) display the worst performance. The interval-dependent tests have the greatest power (Tests 3, 6 and 12 to 15) which is roughly equal to 35% when the sample size is 200, and 95% when $n = 1000$. GT rank, GT KM, Breslow's (rank score) and Lin's tests with weights by Schemper et al. (2009), exhibit similar performance with a loss of power of about 5% in each case. In general, the extension of the follow-up period by 2 time units does not seem to offer much gain (Figure 4.4). Analogous conclusions are drawn about the third case of crossing hazards examined (see Table A.5, Figure A.5).

Figure 4.4: Power of 18 tests for proportional hazards, for different sample sizes and cut points for the scenario of crossing hazards with initial HR = 0.65 and subsequent HR = 1.10.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	27.4	85.1	32.4	87.0
2	17.2	67.0	26.6	83.7
3	22.9	77.7	35.8	95.1
4	27.4	85.1	32.3	87.5
5	19.9	68.4	28.4	83.9
6	22.4	78.1	33.3	95.1
7	24.7	82.8	34.5	92.3
8	24.6	82.9	34.6	92.3
9	7.4	53.8	32.1	90.4
10	25.4	82.7	35.8	92.2
11	10.0	78.9	19.7	84.9
12	23.1	77.8	35.3	95.1
13	22.9	77.8	34.6	95.0
14	23.7	79.3	36.3	95.8
15	23.1	77.8	35.3	95.1
16	24.1	82.6	34.6	92.3
17	27.6	85.1	33.8	88.9
18	18.7	66.5	26.7	82.8

Table 4.5: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the scenario of crossing hazards with initial HR = 1.10 and subsequent HR = 0.65, for different sample sizes n and cut points 2 and 4.

When the initial HR is equal to 1.10 and the subsequent is 0.65, the results are somewhat different. Once again, Gill & Schumacher's and Lin's test (with Xu & O'Quigley's weights) display the worst performance in general. When the $t_{CP} = 2$, GT test with $g(t) = t$, the equivalent test by Cox, and Breslow's cumulative hazards score test are considered the optimal choices. Nevertheless, when $t_{CP} = 4$, the interval-dependent tests reach their crescendo, outperforming all the others. Engagingly, Tests 7,8, 10 and 16 are close in terms of power to the best choices, whether $t_{CP} = 2$ or $t_{CP} = 4$ (Table 4.5, Figure 4.5). The same holds for the fourth case of crossing hazards investigated in this thesis (see Table A.6, Figure A.6).

Long-term Survivors

Finally, the findings for the long-term survivors with initial HR = 0.65 and final HR = 0.65^2 are presented in Table 4.6 and Figure 4.6. The interval dependent tests along with Lin's proposal with weights by Schemper et al. (2009) are empirically the most promising options, while the other choice of weights for Lin's test and Gill

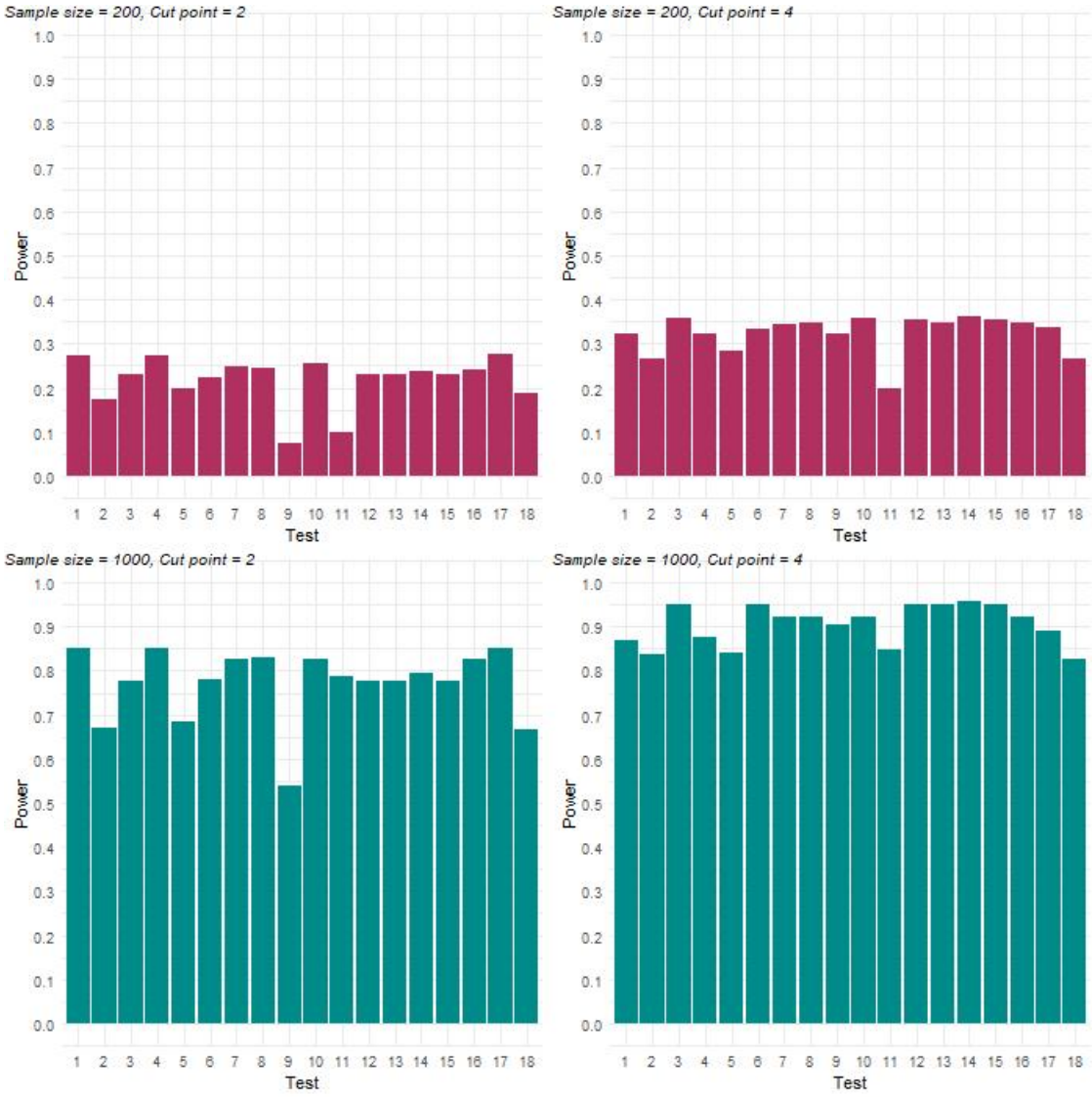

Figure 4.5: Power of 18 tests for proportional hazards, for different sample sizes and cut points for the scenario of crossing hazards with initial HR = 1.10 and subsequent HR = 0.65.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	18.9	65.6	24.9	66.1
2	12.9	51.2	20.5	66.2
3	18.3	64.8	27.5	85.5
4	18.9	65.4	23.2	68.1
5	14.2	51.3	20.8	66.4
6	17.6	65.7	26.5	85.3
7	18.8	66.4	27.5	77.6
8	18.8	66.6	27.5	77.6
9	1.7	11.9	22.6	71.1
10	22.8	67.6	29.1	78.1
11	1.4	11.7	1.7	49.0
12	18.4	65.0	26.6	85.2
13	18.2	64.8	25.6	84.8
14	19.7	68.1	27.5	85.9
15	18.4	65.0	26.6	85.2
16	18.9	66.7	27.5	77.6
17	19.7	67.5	25.0	71.8
18	14.4	50.6	20.6	64.4

Table 4.6: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the scenario of long-term survivors with initial HR = 0.65 and subsequent HR = 0.65^2, for different sample sizes n and cut points 2 and 4.

& Schumacher's approach display a severe lack of power. Once again, Tests 7,8 and 16 approximate the performance of the optimal tests in each case. Similar patterns are observed for the second scenario of long-terms survivors (see Table A.7, Figure A.7).

Despite the fact that there is not a unique test which outperforms the others under all non-PH patterns and special scenarios examined in this thesis, it has been shown that three out of the 18 are close to the optimal option with a usual loss of power of about 5%. These are the rank and KM tests from the GT family (the second is the default transformation in the function cox.zph from the well-known package survival in R), along with Lin's proposal using the weights of Schemper et al. (2009). While the GT tests are famous amongst statisticians, Lin's test is an innovative method for checking the proportionality assumption, which also provides weighted HR estimates, suitable for subsequent analysis and interpretation of the nature of the data (see Chapter 5 for more).

Figure 4.6: Power of 18 tests for proportional hazards, for different sample sizes and cut points for the scenario of long-term survivors with initial HR = 0.65 and subsequent HR = 0.65^2.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 5

Tests for treatment effect

5.1 Estimating treatment effect under non-PH

The traditional log-rank test and the conventional Cox PH model are habitually used for the analysis of trials with time-to-event endpoints. As previously indicated, both methods achieve maximum power and estimation accuracy under the condition of proportional hazards. When a non-PH pattern is observed their power is reduced and the Cox model's HR estimate is severely biased. For instance, if a delayed treatment benefit is detected, the estimated effect will be diluted, as the PH model produces an average HR across the total follow-up time, misleading the investigators involved in the trial. At the same time, even if the above HR is reported as an average, this estimate has been proved to be dependent on the censoring distribution and thus, the estimated effect turns out to be trial-specific (Boyd et al., 2012; Nguyen & Gillen, 2012).

Numerous alternative summary measures have been proposed in the literature, with weighted Cox HRs and median or Restricted Mean Survival time (RMST) difference between arms being the most prevalent. Weighted HRs can be estimated after implementing the max-combo test or any other variant of the log-rank test (see sections 5.2 and 5.3). Further appealing methods have also been suggested, such as the weighted HR estimates proposed by Xu & O'Quigley (2000) or Boyd et al. (2012). These approaches have been praised in a relatively recent paper by Rufibach (2019), as they provide robust estimators against the censoring distribution, which at the same time, are equal to the unweighted HR when the proportionality assumption is valid. Usually, difficulties in clinical interpretation of a unique treatment effect measure when in fact the effect is time-dependent, resulted in reporting the RMST difference along with the corresponding confidence interval (see section 5.4). Other choices involve the estimation of a time-varying treatment effect or the

combined reporting of the Cox model's estimates and some weighted counterparts. The latter approaches give a detailed knowledge of the history of the trial but are more complicated in terms of interpretation. In a similar spirit, one can also report piecewise HRs, if there is evidence that there is a point in time where the treatment effect displays a different pattern than before. Of course, the choice of the summary measure affects both the analysis and the design of the clinical trial and should be carefully considered.

In clinical trials, most of the time, the interest is focused on comparing two treatments, i.e., two groups of patients who follow different therapeutic approaches. So before even a measure for treatment effect is reported, it is important to test if the two therapies differ significantly. In the next sections, the theoretical basis of several testing methods for this cause will be presented, along with their corresponding measures, if there are any. These methods will also be examined and compared via simulations in the next chapter.

5.2 Weighted log-rank tests and variants

5.2.1 The Fleming-Harrington family

In Chapter 2, section 2.3, the traditional log-rank test for the comparison of two survival curves was thoroughly explored. In a study where m events have been observed, one can briefly say that the log-rank test arises from the combination of m 2×2 contingency tables which display group versus survival status at each failure time. Considering also that, under the null hypothesis of no difference between the survival profiles of the populations of interest, the number of events in the first group follows a hypergeometric distribution, a statistic of the form presented in (2.5) occurs, or

$$X^2_{LR} = \frac{\left[\sum_{j=1}^{m}(d_{1j} - e_{1j})\right]^2}{\sum_{j=1}^{m} \frac{d_j r_{1j} r_{2j}(r_j - d_j)}{r_j^2(r_j - 1)}} \tag{5.1}$$

where

- d_j is the number of events taking place at t_j,
- r_j is the number of subjects at risk at t_j,
- d_{1j} is the number of events taking place at t_j in group 1,
- $e_{1j} = d_j \frac{r_{1j}}{r_j}$ is the expected number of events at t_j in group 1, and lastly,

- r_{1j} and r_{2j} are the number of subjects at risk in groups 1 and 2, respectively.

Under H_0, X^2_{LR} has an asymptotic chi-square distribution with 1 degree of freedom.

Despite the simplicity and usefulness of the log-rank test, a statistician should always bear in mind that it is the optimal choice when the assumption of PH holds, but its power is diminished as the hazard ratio deviates more and more from a constant function. Of course, it is performed in almost every analysis of survival data where the comparison of groups is under the microscope. It is valid even when a non-PH pattern is observed and theoretically the most powerful test when the hazards are proportional.

To increase its power under non-PH, Fleming & Harrington (1982) proposed a variation called the *weighted log-rank test*. The traditional log-rank test's incompetence to detect important differences between survival curves which occur either at the beginning or at the end of the follow-up time, motivated the two professors of Biostatistics to think of an alternative approach for comparing two survival functions. They initially introduced the G^ρ family of statistics, where the weighting functions are defined as follows:

$$w(t) = \{\hat{S}(t)\}^\rho, \rho \geq 0, \tag{5.2}$$

and $\hat{S}(t)$ is the KM estimate[1] of the survival function based on the whole dataset. More specifically, the statistic in (5.1) is modified, and is now given by

$$X^2_w = \frac{\left[\sum_{j=1}^m w_j(d_{1j} - e_{1j})\right]^2}{\sum_{j=1}^m w_j^2 \frac{d_j r_{1j} r_{2j}(r_j - d_j)}{r_j^2(r_j - 1)}} \tag{5.3}$$

where $w_j = \{\hat{S}(t_j)\}^\rho$ and the rest quantities are defined like before. Again, under the assumption that the distributions of survival times in the two groups are identical, X^2_w asymptotically follows a chi-square distribution with one degree of freedom.

This idea improved the power of the comparison test in situations where early differences occurred. Undoubtedly, this approach turned out to be useful on occasions where, for instance, a treatment reduced the hazard for some initial period, but its effect on the hazard decreased later on. This change is justified by the fact that the family of weights given by (5.2) consists of decreasing functions since the survival curve is always decreasing itself. As a result, the beginning of the follow-up period is more definitive for the outcome of the comparison than middle or the end of the study. Of course, for $\rho = 0$, (5.3) corresponds to the traditional log-rank test,

[1] Usually the left continuous version of the KM estimator is used.

while for $\rho = 1$, it seems that the class contains as a special case a test essentially equivalent to Peto & Peto's generalization of the Wilcoxon test (Peto & Peto, 1972). As expected, the latter is sensitive to early differences in survival between groups, since $w_j = \hat{S}(t_j)$.

Undoubtedly, this family of FH tests generalized and enhanced the power of the simple log-rank test, however, it could not adequately detect differences between treatment arms for which the survival curves did not separate until a certain interval of time has elapsed[2]. That being the case, Fleming & Harrington (1991) extended this definition to the $G^{\rho,\gamma}$ family of statistics, with weights defined as

$$w(t) = \{\hat{S}(t)\}^{\rho}\{1 - \hat{S}(t)\}^{\gamma}, \tag{5.4}$$

for $\rho \geq 0$, and $\gamma \geq 0$. In contrast to the previous definition, these weighting functions give more flexibility regarding the choice of the most influential time interval for the test statistic. When $\gamma = 0$ in equation 5.4, the $G^{\rho,\gamma}$ class of statistics reduces to the G^{ρ} family, placing more weight on earlier events. When only $\rho = 0$, more weight is given to later events. If $\rho = \gamma$, the test is more powerful for differences in the middle of the total follow-up time. Of course, when $\rho = \gamma = 0$, the FH test is equivalent to the unweighted log-rank test.

Apart from the FH family of tests, other types of weighted log-rank tests have also been proposed in the literature. The most popular amongst them are the Gehan-Wilcoxon, the Tarone-Ware and the Modified Peto-Peto test.

Gehan-Wilcoxon (or Generalized Wilcoxon) test

The Gehan-Wilcoxon test uses the number of individuals at risk r_j at time t_j as the weight; thus, in equation 5.3, $w_j = r_j$. Since the weight is the number of individuals at risk, the Gehan-Wilcoxon test places more emphasis on the information at the beginning of the survival curve, where the number at risk is larger, allowing early failures to receive more weight than later events. It has been proved to be a powerful test even when the PH assumption does not hold (Gehan, 1965; Karadeniz & Ercan, 2017).

Tarone-Ware test

The Tarone-Ware test places more weight on hazards in the early periods, just as the Gehan-Wilcoxon test does. More precisely, it uses the square root of the

[2]This is essentially the case of a delayed effect, discussed in section 3.1.

number of individuals at risk at each failure time as weights, i.e., $w_j = \sqrt{r_j}$ for $j = 1, 2, \ldots, m$ (Tarone & Ware, 1977). Without a doubt, the weight used in the Tarone-Ware test is greater than the weight used in the log-rank test ($w_j = 1\ \forall j$) but less than the weight used in the Gehan-Wilcoxon test.

Modified Peto-Peto test

Finally, the modified Peto-Peto test extends the initial test suggested by Peto & Peto (1972). It places even greater weight on the beginning of the study since $w_j = \tilde{S}(t_j)r_j/(r_j + 1)$. Careful consideration should be given here: despite the fact that the formula of weights includes an estimator of the survival function, in the case of Peto-Peto's test and its modified version, a different estimate than the one produced by the KM method is typically preferred (Karadeniz & Ercan, 2017).

5.2.2 Versatile weighted log-rank tests

Even though an appropriate choice of ρ and γ in the extended FH family can result in a well-powered test, little is the a-priori knowledge on how and when a significant difference between two curves and a non-proportional hazards pattern can evolve. On many occasions, investigators are unable to predict the shape of the survival functions and even when they approximately do, they cannot specifically define the time point or interval where the difference will be significant, imposing difficulties on the analysis. Thus, which choice of ρ and γ is optimal, especially before a clinical trial is conducted or even designed? To answer this crucial question, a combination of FH tests can be implemented, including multiplicity correction, not only to compare two survival functions, but also to track the time frame in which the difference achieved its greatest magnitude. Throughout the years, many statisticians have considered this option: Lee (1996), Lee (2007), Karrison (2016), and Lin et al. (2020) to name a few. Each one of the proposed approaches employs a combination of weighted log-rank tests mentioned in the previous section and the multiplicity correction is based on the assumption that the vector with elements the individual weighted test statistics follows a multivariate normal distribution (Karrison, 2016). The individual statistics are in fact equal to the square root of the statistic given by (5.3) for some particular choice of weighting function, and they are called the *z-statistics*, since each one follows a univariate standard normal distribution. After implementing one of these tests, it is also possible to obtain weighted parameter estimates from the Cox model, choosing as weights the ones corresponding to the FH test with the smallest p-value (see section 5.2).

Lee (1996)

Lee (1996) evaluated the maximum over four z-statistics, derived from $G^{0,0}$, $G^{2,0}$, $G^{0,2}$, and $G^{2,2}$ tests, as well as their average. The reason why he chose this particular combination is because it simultaneously examines four different scenarios, three of which perform well under dissimilar non-PH patterns. Table 5.1 matches each of the above FH weighted tests with the scenario under which it is expected to have the highest power amongst the others. These comments stem directly from what was discussed about the FH family in section 5.2.1.

(ρ, γ)	$w(t)$	Maximum Power
$(0,0)$	1	Proportional Hazards
$(2,0)$	$[\hat{S}(t)]^2$	Early Effect
$(0,2)$	$[1-\hat{S}(t)]^2$	Late Effect
$(2,2)$	$[\hat{S}(t)]^2[1-\hat{S}(t)]^2$	Middle Difference

Table 5.1: FH tests involved in Lee's (1996) proposal and expected scenarios of optimal performance.

Lee (1996) conducted a simulation study, comparing the individual members of the $G^{\rho,\gamma}$ family involved in his approach and the combined statistics, under the PH scenario and cases of early, middle and late hazard differences. As anticipated, the individual tests performed better than the others under the assumption to which they were matched in Table 1. Interestingly, the combined statistics, i.e., the maximum and their average, were nearly as sensitive as the most powerful individual statistic for detecting a specific local alternative. It also appeared that the maximum of the individual statistics performs slightly better than the average of those statistics.

Lee (2007)

Lee (2007) considered three combination tests based on two z-statistics corresponding to $G^{1,0}$ and $G^{0,1}$ from the FH family. If Z_1 and Z_2 are the z-statistics, he compared their performance with the power of their maximum $\max(|Z_1|, |Z_2|)$, and the statistics $|Z_1 + Z_2|$ and $(|Z_1| + |Z_2|)/2$. Simulation results confirmed that the maximum test nearly maintains the sensitivity of the statistics Z_1 and Z_2 for the corresponding survival differences of early and late effect, and is more versatile than both across several scenarios which are dissimilar to the previous and/or more complex. As for the power of the other proposed statistics, his simulation showed that they do not perform better or at least as well as Z_1 and Z_2.

Karrison (2016)

Karrison (2016) generalized the optimal test suggested by Lee (2007), since he considered $\max(|Z_1|, |Z_2|, |Z_3|)$, where Z_1, Z_2 and Z_3 are z-statistics obtained from $G^{1,0}$, $G^{0,1}$ and $G^{0,0}$ tests. This combination also covers the case of proportional hazards, along with the early and the late effect. Karrison's test maintains the type I error rate and provides increased power in comparison to the log-rank test under early and late difference alternatives; however, $\max(|Z_1|, |Z_2|, |Z_3|)$ is associated with a small to moderate power loss relative to the more optimally chosen test. It is also quite close to $\max(|Z_1|, |Z_2|)$ in terms of performance.

Lin et al. (2020)

Lin et al. (2020) suggested to use the maximum of the absolute values of four FH weighted statistics: those which correspond to $G^{0,0}$, $G^{1,0}$, $G^{0,1}$ and $G^{1,1}$. Consequently, exactly like Lee (1996), the proposed test will provide relatively good coverage across a range of possibilities: proportional hazards, early, middle and late difference configurations. According to the relative paper, the MaxCombo test, as they call it, is robust against various patterns of non-PH and it provides a strong advantage under late effect or crossing hazards, scenarios commonly observed in immuno-oncology. At the same, it achieves acceptable power under early effect and proportional hazards compared to the traditional log-rank test.

5.2.3 Combinations with other tests

The extension of the traditional log-rank test to a weighted version has undoubtedly provided greater flexibility and better properties on occasions where non-PH patterns are present. Nevertheless, this approach just takes into account the information accumulated during a specific period more than other time intervals, based on what has been observed (if the method is directly applied during the analysis) or has been expected (if the method is chosen a priori at the design stage). Even when versatile weighted tests are used, oftentimes the variety of alternative scenarios under consideration is restricted. Of course, various combinations of weighted log-rank tests can be constructed; the ones presented in the previous section are just the most famous in the literature. Statisticians can choose a wide range of values for ρ and γ and employ the multivariate normal distribution of the corresponding statistics to perform a test suitable for their data. As a matter of fact, major statistical software

packages, such as R, already include functions performing these types of versatile tests in accordance with the user's preference.

As mentioned in Chapter 3, section 3.1, non-proportionality takes on many forms. It will be extremely naive to assume that the four non-PH scenarios presented in the current thesis are enough to represent every possibility. Intuitively, each weighted counterpart of the log-rank test corresponds to a particular non-PH pattern, and it is difficult to predict beforehand the shape of the survival curves and the relationship connecting them. In other words, it is hard to find the optimal (versatile) weighted test. To overcome this problem, different approaches have been developed, combining the great power of the log-rank test when proportionality holds with some of the tests for the PH assumption discussed in Chapter 3.

To understand better why such an approach would work, it is important to remember that when the hazards of two groups are proportional, i.e., $h_1(t)/h_2(t) = \theta$, then

$$S_1(t) = [S_2(t)]^\theta, \tag{5.5}$$

for some $\theta > 0$. If $\theta = 1$ the survival profiles in the two populations are identical. Consequently, the equivalence of the survival functions, and thus, the absence of treatment effect, is a special case of the proportionality of hazards. When the PH assumption is invalid, the same holds for equation 5.5. This means that

$$S_1(t) \neq [S_2(t)]^\theta, \forall \theta \in (0, +\infty).$$

Now, imagine performing a test that utilizes both the log-rank test and a testing procedure for the PH assumption. If there is indeed a significant treatment effect and proportionality holds, the log-rank test has the maximum power to reject the null hypothesis of no difference between the two arms. On the other hand, when the PH assumption does not hold, the same is true for the null hypothesis, so the proportionality test should be able to provide evidence against it. Based on this rationale, several tests have been proposed in the literature. Two of them are presented below.

Breslow combo test (Breslow et al., 1984)

Recall Breslow, Elder, and Berger's proposal for testing the PH assumption: for the two-sample case, they suggested an alternative model given by (3.15). The null hypothesis of proportional hazards, i.e., $\gamma = 0$, versus the alternative that $\gamma \neq 0$, was then tested by performing a score test for the unknown parameter γ, and resulted

in a test statistic X_2^2, which corresponds to (3.17). When the PH assumption holds, it follows a chi-square distribution with 1 degree of freedom.

According to Breslow et al. (1984), if one wants to test another null hypothesis H_0, i.e., that both β and γ in (3.15) are equal to zero, versus the alternative that only $\gamma = 0$, then the log-rank test would be derived from the corresponding score test. Using the same notation as the one in section 3.2.3, the score statistic for the log-rank test would be

$$X_1^2 = \frac{[\sum_{j=1}^{m} d_{1j} - p_j(0,0)]^2}{\sum_{j=1}^{m} p_j(0,0)q_j(0,0)} = \frac{[\sum_{j=1}^{m} d_{1j} - r_{1j}/r_j]^2}{\sum_{j=1}^{m} r_{1j}r_{2j}/r_j^2} \tag{5.6}$$

where $r_j = r_{1j} + r_{2j}$ is the total number of subjects at risk at time t_j and $p_j(0,0)$ is calculated as in (3.16). Once again, X_1^2 follows a chi-square distribution with 1 degree of freedom. Notice that this is equivalent to the statistic in (2.5) when there are no ties.

Therefore, in order to test the null hypothesis

$$H_0 : \beta = \gamma = 0$$

versus a more generic alternative

$$H_1 : \beta \neq 0 \text{ or } \gamma \neq 0,$$

Breslow et al. (1984) suggested using both statistics X_1^2 and X_2^2. It was shown that under H_0, as $n \to \infty$, the statistics X_1 and X_2 have independent normal distributions. Therefore, when the log-rank test and acceleration test are simultaneously implemented, a multiplicity correction to avoid an inflated type I error can be done by using a maximum-modulus test based on $\max(|X_1|, |X_2|)$. Another option is to use $X_1^2 + X_2^2$, which under H_0 follows a chi-square distribution with 2 d.f. as $n \to \infty$. Lin et al. (2020), praised this particular approach since it seems to result in a potential power gain under crossing hazards.

Joint test by Royston & Parmar (2014)

The *joint test*, as Royston & Parmar (2014) called it, is a combination of the log-rank test and a test from the GT family. Despite the fact that any test can be used, the authors presented their findings using as time function $g(t)$ the ranks of the event times, i.e., they preferred a proportionality test based on the correlation between scaled Schoenfeld residuals and the ranks of the failure times.

This idea was mainly developed to overcome the problem of an inflated type I error which is the result of the following common analysis approach: In most cases, the sample size of a trial is calculated via a log-rank test. However, due to new therapies and breakthroughs in the field of medicine, non-PH patterns are more frequent than ever. Naturally, the traditional log-rank test is not the optimal test since its power is reduced under non-proportionality. At the same time, when the data are collected, a test for PH must be performed to validate the results occurring from usual analysis techniques such as the fit of a Cox PH model. An issue that arises here, is that the log-rank test and a subsequent test from the famous GT family double the probability of a type I error.

To rectify this issue, Royston & Parmar (2014) brainstormed the next idea: Under the null hypothesis of identical survival profiles in the two groups of interest, it holds that the log-rank and the GT test have independent corresponding statistics, each of which follows a chi-square distribution with 1 d.f. and thus, their sum has an asymptotic chi-square distribution with 2 degrees of freedom. Consequently, this joint test can be utilized to simultaneously check the proportionality assumption and find evidence in favor of a significant treatment effect.

Even though their approach is not suggested for use routinely, it has proved to be quite powerful under increasing or decreasing HR, outperforming the log-rank test. Of course, the latter has greater power than the joint test under PH, but they are still close. Keep in mind that the joint test has been mainly presented, in the literature, as an alternative approach for the calculation of the sample size during the design of a trial, and not as an analysis procedure.

5.3 Cox regression under non-PH and related models

The conventional Cox PH model, presented in section 2.4, has proved to be one of the main statistical tools used in survival analysis. It offers great flexibility in comparison to parametric models which assume a specific form of the baseline function, while simultaneously adjusting for the effect of many covariates. Its only "restrictive" property is the assumption of proportional hazards: in section 2.4.3, it was shown that, if Cox regression is used for modeling the data, then for any two individuals the ratio of the corresponding hazard functions is independent of time. In reality, however, this is rarely the case.

When hazards are indeed proportional, the Cox model yields unbiased, easily

interpretable estimates and the power of the corresponding tests for significance is at its highest possible level. Such good properties oftentimes lead statisticians to ignore the problem of non-proportionality, especially when the deviation from PH is small. Some claim that the implementation of the standard Cox model is an acceptable approach under non-proportionality as long as there is not an inversion of effect for the covariates, i.e., the sign of the log-HR does not change over time. They also propose to interpret the estimates as average HRs. Nevertheless, when the fundamental assumption of the Cox model does not hold, it is impossible to gain good enough estimates for the treatment effect or the effect of any other covariate. Apart from the bias, it also seems that the results obtained in this way are study-specific since they are sensitive to the censoring pattern of the data at hand. No reliable inference can be drawn from such an analysis.

5.3.1 An intuitive interpretation of the standard HR estimate under non-PH

Before proceeding to the presentation of alternative methods, an intriguing question here is why some investigators consider the estimate of the HR provided by Cox model as an average? Is it truly an average value of the underlying time-varying HR and what does that mean? To examine this statement, a simulation can be used. For simplicity, the two-sample case is considered. The distribution of survival time within each population has been selected to be a piecewise exponential, with initial rate equal to $\lambda_0 = \lambda_0' = 1$ for both samples. After a certain time point τ_1 (here $\tau_1 = 0.5$) both rates change: for the first group $\lambda_1 = 0.5$, while for the second $\lambda_1' = 0.3$ after τ_1. Figure 5.1 illustrates the survival curves in the two samples. Each group includes 1000 subjects and 1000 repetitions are implemented in total. For each repetition, the Cox PH model is fitted to the dataset applying administrative censoring at different time points consisting a sequence of the form $t_k = \tau_1 + 0.1 \cdot k$, $k = 0, 1, \ldots, 55$. The HR occurring as a MPLE is therefore saved for each repetition and each choice of k. Then, the mean value of all HRs is computed for each time point t_k and the results are graphically displayed in Figure 5.2. It is evident that the time point at which the study ends determines the value of the HR estimate given by the Cox model. In this particular example, it ranges from roughly 0.75 to a value a little greater than 1. Notice that as the follow-up period is extended the estimate decreases. This comes naturally as a result of the nature of the data: up to $\tau_1 = 0.5$ the real HR is equal to one but as time passes, the survival profile of the second group is better than that of group 1. Consequently, the HR estimate for

the second group compared with the first becomes smaller and smaller as the study period is prolonged.

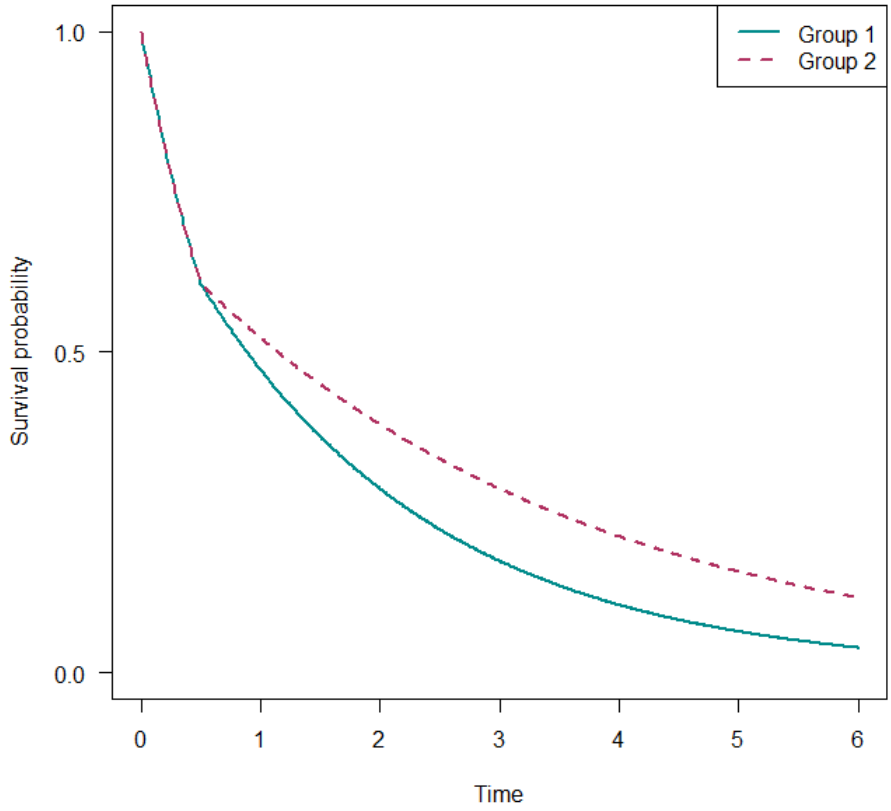

Figure 5.1: Survival functions of two populations for which the survival time distribution is piecewise exponential, with initial hazard rates $\lambda_0 = \lambda_0' = 1$ before $\tau_1 = 0.5$, and rates $\lambda_1 = 0.5$ and $\lambda_1' = 0.3$ after τ_1 for groups 1 and 2, respectively.

This change in the HR gave rise to the idea that the Cox model's estimate under non-PH is, in a sense, an average of the real HR. Note that the real HR here is a piecewise constant function, i.e.,

$$\mathrm{HR}(t) = \begin{cases} \mathrm{HR}_1 = 1, & \text{if } t \leq 0.5 \\ \mathrm{HR}_2 = 0.6, & \text{if } t > 0.5 \end{cases}$$

Therefore, if the interpretation of the estimate as an average HR is correct, then it should be equal to a weighted mean of HR_1 and HR_2. Three possible weights are being explored, based on

1. the percentage of time spent in the internals $[0, 0.5]$ and $(0.5, t_k]$,
2. the expected number of events in each of the internals $[0, 0.5]$ and $(0.5, t_k]$, and finally,

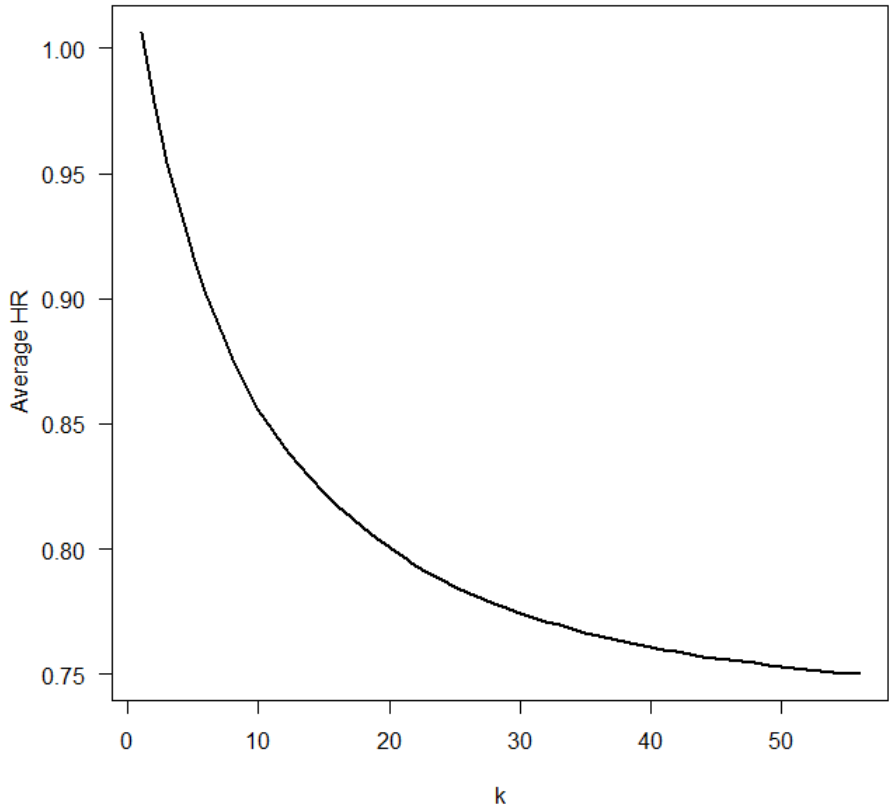

Figure 5.2: Average hazard ratio in the interval $[0, t_k]$, with $t_k = 0.5 + 0.1 \cdot k$, versus $k = 0, 1, \ldots, 55$.

3. the cumulative hazard for a randomly chosen individual.

First approach

Suppose that the follow-up period ends at a time point $t_k, k = 0, 1, \ldots, 55$. Then, the percentage of time spent in $[0, 0.5]$ is given by $p_1 = 0.5/t_k$ and the percentage of time spent in $(0.5, t_k]$ is given by $p_2 = (t_k - 0.5)/t_k = 1–p_1$. These weights can be used to calculate either a weighted arithmetic (AM) or a geometric mean (GM) of HR_1 and HR_2, i.e.,

$$\text{AM}_1 = p_1 \cdot \text{HR}_1 + p_2 \cdot \text{HR}_2$$

and

$$\text{GM}_1 = \text{HR}_1^{p_1} \cdot \text{HR}_2^{p_2}.$$

Of course, for $k = 0 \Rightarrow t_0 = 0.5$ and thus $p_1 = 1$ and $p_2 = 0$. This means that if the study ends after 0.5 time units of follow-up, $\text{AM}_1 = \text{GM}_1 = \text{HR}_1$. So, the results up to $t_0 = 0.5$ must represent the case of PH with HR = 1. Figure 5.3 shows that

AM_1 and GM_1 are far from what is given as a MPLE from the Cox model. In other words, these weights do not yield the desired result.

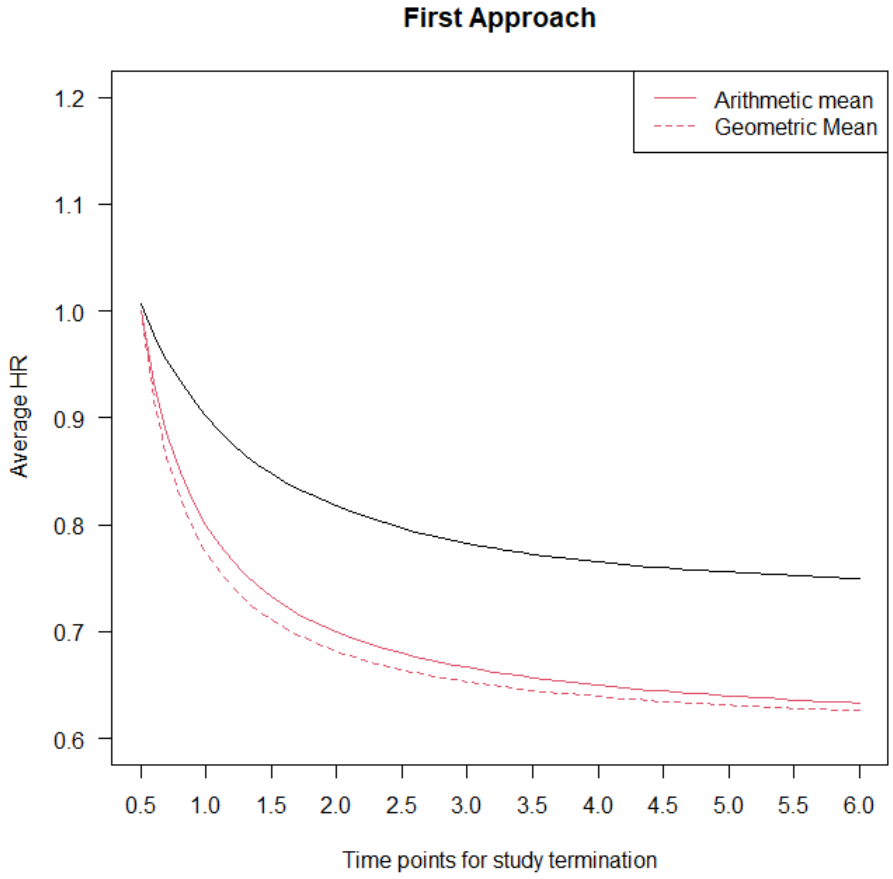

Figure 5.3: Arithmetic and geometric mean of $HR_1 = 1$ and $HR_2 = 0.6$ with weights equal to the percentages of time spent in each of the time internals $[0, 0.5]$ and $(0.5, t_k]$, for $k = 0, 1, \ldots, 55$. The black line corresponds to the HR estimate from the Cox model.

Second approach

Denote by $F_{G_1}(t)$ and $F_{G_2}(t)$ the cumulative distribution functions of groups 1 and 2, respectively. It holds that

$$F_{G_1}(t) = \begin{cases} 1 - e^{-t}, & if\ t \leq 0.5 \\ 1 - e^{-0.5t-0.25}, & if\ t > 0.5 \end{cases}$$

and

$$F_{G_2}(t) = \begin{cases} 1 - e^{-t}, & if\ t \leq 0.5 \\ 1 - e^{-0.3t-0.35}, & if\ t > 0.5 \end{cases}$$

If $F(t)$ is the cumulative distribution function for a random subject in the study, then

$$F(t) = \mathrm{P}(\text{subject in group 1})F_{G_1}(t) + \mathrm{P}(\text{subject in group 2})F_{G_2}(t),$$

and thus,

$$F(t) = \begin{cases} 1 - e^{-t}, & \text{if } t \leq 0.5 \\ 0.5 \cdot [1 - e^{-0.5t-0.25}] + 0.5 \cdot [1 - e^{-0.3t-0.35}], & \text{if } t > 0.5 \end{cases}$$

since the number of individuals belonging to group 1 is equal to the number of individuals belonging to the second group.

The whole study includes $n = 2000$ subjects. The number of events up to a specific time point t_k is a binomial random variable with probability of success[3] equal to $F(t_k)$. Similarly, the number of events taking place in a time interval $(t_{k_1}, t_{k_2}]$ is a binomial random variable with success probability equal to $F(t_{k_2}) - F(t_{k_1})$. Consequently, if the weights are defined as

$$e_1 = \frac{\text{Expected number of events within } [0, 0.5]}{\text{Expected number of events within } [0, t_k]}$$

and

$$e_2 = \frac{\text{Expected number of events within } (0.5, t_k]}{\text{Expected number of events within } [0, t_k]}$$

or

$$e_1 = \frac{n \cdot F(0.5)}{n \cdot F(t_k)} = \frac{F(0.5)}{F(t_k)}$$

and

$$e_2 = \frac{n \cdot [F(t_k) - F(0.5)]}{n \cdot F(t_k)} = \frac{[F(t_k) - F(0.5)]}{F(t_k)},$$

for $k = 0, 1, \ldots, 55$, then an arithmetic and a geometric mean can be constructed as follows:

$$\mathrm{AM}_2 = e_1 \cdot \mathrm{HR}_1 + e_2 \cdot \mathrm{HR}_2,$$

and

$$\mathrm{GM}_2 = \mathrm{HR}_1^{e_1} \cdot \mathrm{HR}_2^{e_2}.$$

Figure 5.4 displays how these two values change in relation to the total follow-up period. It is evident that GM_2 is almost identical to the average HR calculated from the Cox PH model.

[3]Success = Event by the time t_k, for $k = 0, 1, \cdots, 55$.

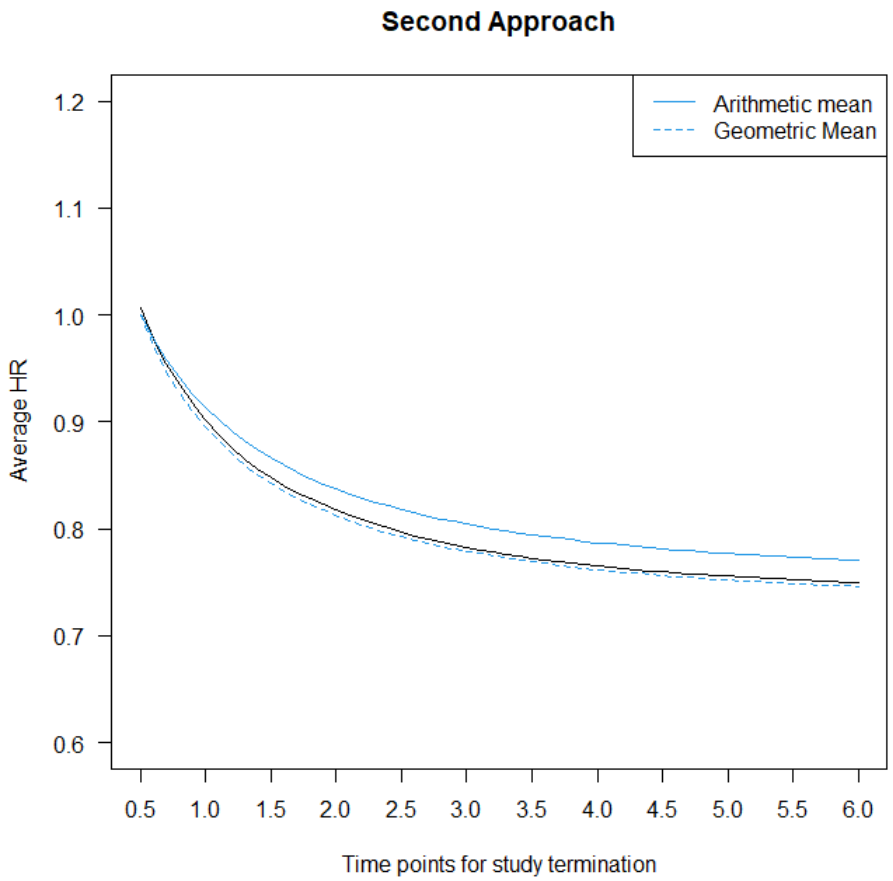

Figure 5.4: Arithmetic and geometric mean of $\text{HR}_1 = 1$ and $\text{HR}_2 = 0.6$ with weights depending on the expected number of events within each of the time internals $[0, 0.5]$ and $(0.5, t_k]$, for $k = 0, 1, \ldots, 55$. The black line corresponds to the HR estimate from the Cox model.

Third approach

Let $\Lambda(t)$ be the cumulative hazard for a randomly selected individual. It holds that

$$\Lambda(t) = -\ln\left[1 - F(t)\right] = \begin{cases} t, & \text{if } t \leq 0.5 \\ -\ln\left[0.5 \cdot (e^{-0.5t-0.25} + e^{-0.3t-0.35})\right], & \text{if } t > 0.5 \end{cases}$$

Choosing as weights

$$c_1 = \frac{\Lambda(0.5)}{\Lambda(t_k)}$$

and

$$c_2 = 1 - c_1 = \frac{\Lambda(t_k) - \Lambda(0.5)}{\Lambda(t_k)}$$

the following arithmetic and geometric mean occur:

$$\mathrm{AM}_3 = c_1 \cdot \mathrm{HR}_1 + c_2 \cdot \mathrm{HR}_2$$

and

$$\mathrm{GM}_3 = \mathrm{HR}_1^{c_1} \cdot \mathrm{HR}_2^{c_2}.$$

Figure 5.5 shows the relationship between each mean and the average HR. It is obvious that this method does not approximate the average HR as well as the second approach. In fact, all methods are compared in Figure 5.6: only the geometric mean of HR_1 and HR_2 with weights the expected number of deaths within each time interval is close to the Cox model's estimate.

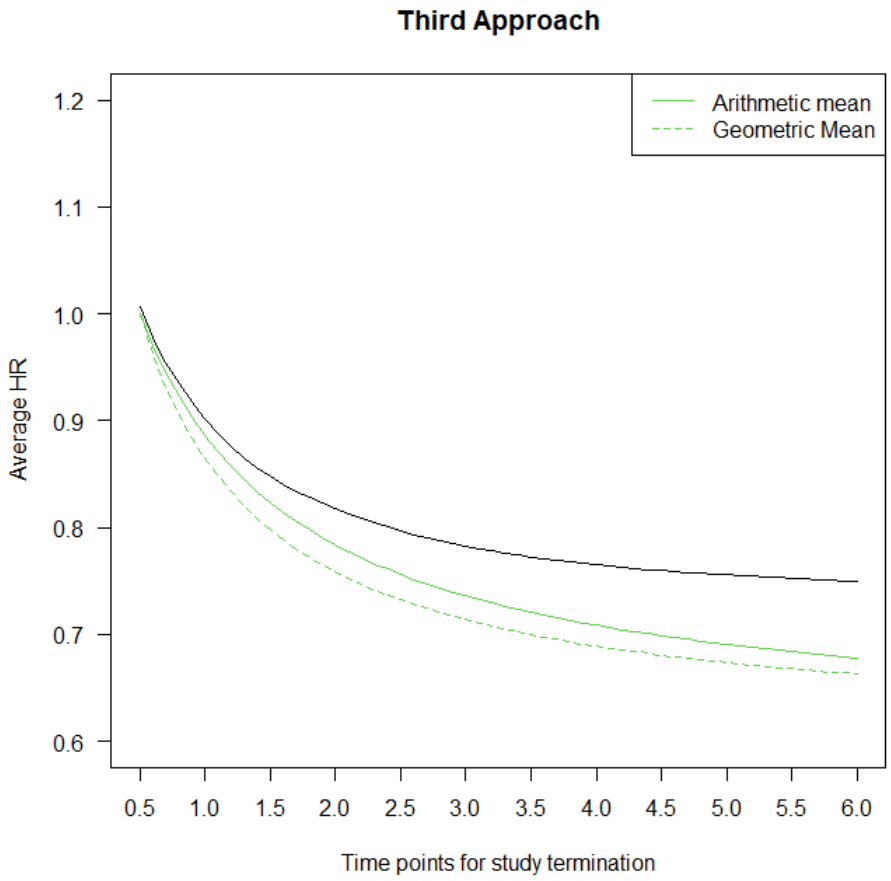

Figure 5.5: Arithmetic and geometric mean of $\mathrm{HR}_1 = 1$ and $\mathrm{HR}_2 = 0.6$ with weights depending on the cumulative hazard of a randomly selected individual. The black line corresponds to the HR estimate from the Cox model.

Now remember what was discussed in Chapter 2 for the piecewise exponential distribution. It was mentioned that it is a very useful distribution, particularly for the simulation of survival data, since it is quite flexible and an appropriate choice

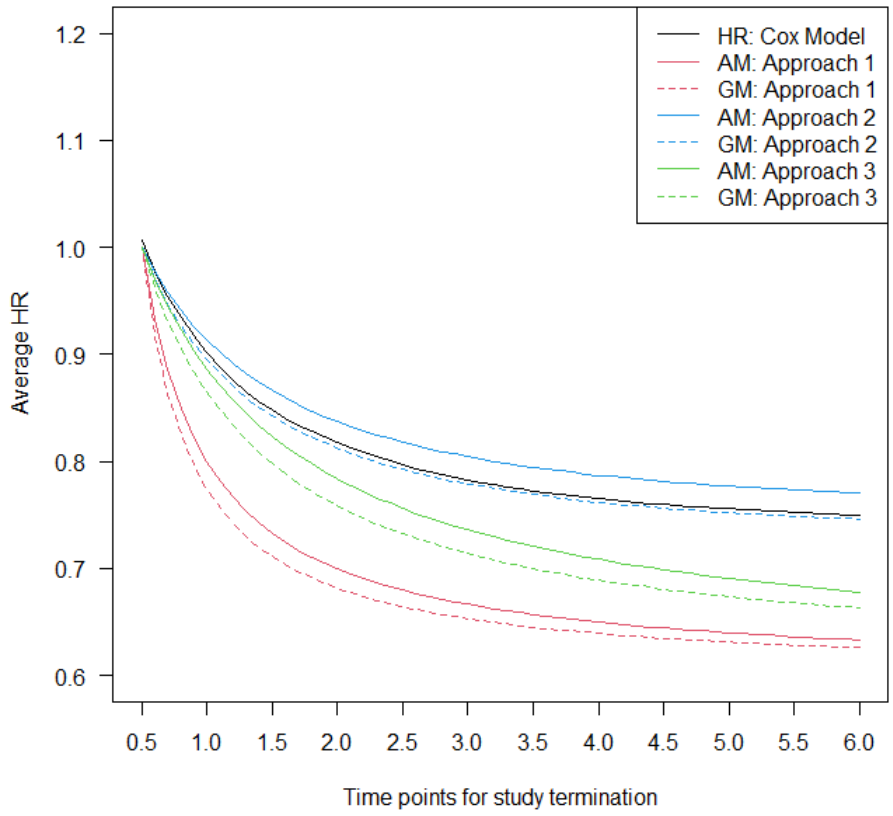

Figure 5.6: Arithmetic and geometric mean of $HR_1 = 1$ and $HR_2 = 0.6$ using three different approaches, compared with the average HR estimator of the Cox model.

of time intervals and hazard rates could mimic the behavior of any real-life dataset. Intuitively, any survival profile can be thought to stem from a piecewise exponential distribution if the time intervals are narrow enough. Consequently, in combination with the findings of the current section, one can interpret the Cox model's HR as a geometric mean of the individual constant HRs, using as weights the proportion of the expected number of events within each time interval per total expected number of events.

5.3.2 Cox model modifications and alternative estimates for the HR under non-PH

Since, under non-proportional hazards, the Cox model's HR estimate underestimates the real HR for some periods of time, and overestimates it for others, it is not quite efficient to base the inference on this particular approach. Alternative methods for the analysis of non-proportional data have been proposed throughout the years, which, in a sense, extend the conventional Cox PH model. The most

popular amongst them, are:

1. *The stratified Cox model*: Suppose that the assumption of PH has been rejected for one or more covariates. Then, a new model can be fitted to the data, assuming that the baseline hazard is different for each level[4] of the variable violating the PH condition. For simplicity, suppose that the proportionality of hazards is invalid only for one covariate with k categories. Then the stratified Cox model is given by

$$\lambda_{i,s}(t) = \lambda_{0s}(t)\exp(\beta' x_i),\ \ s = 1, 2, \ldots, k.$$

   As usual, β is estimated from the corresponding partial likelihood. Notice that the estimate of β is independent of the category s of the predictor being stratified. Therefore, the effect of any other covariate is assumed to be the same across all strata. A drawback of this method is that the effect of the covariate based on which the stratification was implemented, cannot be computed. If this covariate is an indicator for the type of treatment received by a subject, then it would be impossible to estimate the treatment effect using the stratified Cox model. Thus, it is preferable to avoid employing this model when the variable of interest is the one not satisfying the PH assumption. Also, it is suggested that the number of strata should be small, otherwise the complexity of the model would increase unnecessarily.

2. *The extended Cox model*: It is the model presented in (3.25). The HRs for each covariate are given as functions of time. It is a useful approach if the investigators desire to predict the survival profile of the subjects involved in a study. Nevertheless, it does not always provide a clear-cut answer to questions, such as "Which treatment is better?", "Are patients with a specific characteristic A doing better than patients with a specific characteristic B?", etc.

3. *The change-point Cox regression*: It is essentially a special case of the extended Cox model. Sometimes, there may be indications that the hazard ratio is constant within specified time intervals. Other times, this model is just used because proportionality is invalid, but a simpler approach than the ones mentioned above is preferable. However, such an analysis is based on the assumptions of constant HRs within each period and a sudden change at the

[4]If the variable is not qualitative, but quantitative, its range should be split into categories. If more than one covariates are violating the PH assumption, a combination of their categories should be used for the determination of the model.

cutpoint between two (or more) periods. This assumption is quite unrealistic and again, piecewise constant HRs cannot always result in straightforward answers about the superiority of a therapy over another. They, however, offer a better insight into the history of the trial than a single summary measure.

4. *The weighted Cox regression*: It is similar to the simple Cox model, providing one measure of relative risk for each covariate. It has been proved to be efficient for small samples and also to yield more robust estimates than the traditional Cox model under the presence of censoring. The estimates of the HRs are computed solving the partial likelihood score equations after some weights are introduced for each subject and/or covariate.

The burning question of this chapter is how to test the significance of the treatment effect. The stratified model does not allow for the estimation of a treatment effect when the PH assumption is violated for the corresponding variable, and the extended Cox model is more informative of the history of the study rather than the effect of the included covariates. Of course, the addition of a time-varying parameter for the treatment indicator may provide interesting results but the options regarding the form of the time-function for the HR are endless. Therefore, the interest in this section is focused only on the weighted counterpart of the Cox PH model and an interesting test for treatment effect based on the combination of multiple single change-point regression models.

5.3.3 Weighted Cox regression

In order to gain weighted estimates for the HR, one should solve a modified version of the partial likelihood score equations, defined by (2.9) and (2.10). More specifically, the system of equations takes the form

$$\frac{\partial \ell_w(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i w_j(t_i) \left[x_{ij} - \frac{\sum_{\ell \in R_i} x_{\ell j} \exp(\beta' x_\ell)}{\sum_{\ell \in R_i} \exp(\beta' x_\ell)} \right] = 0 \tag{5.7}$$

for $j = 1, 2, \ldots, p$, where $w_j(t)$ is a weighting function, and all other quantities are defined as in section 2.4.1. In equation (5.7) a weighting function $w_j(t_i)$ which permits the contributions to β at each failure time to be weighted differently has been introduced. In the standard Cox model analysis, $w_j(t_i)$ is always 1. As for the information matrix, it is obtained, as usually, by taking minus the second derivatives of the partial log-likelihood introducing the appropriate weights. Thus, the (j, k)

entry of the information matrix is given by the formula

$$I_{jk}(\beta) = \sum_{i=1}^{n} \delta_i w_j(t_i) w_k(t_i) \left[\frac{\sum_{\ell \in R_i} x_{\ell j} x_{\ell k} \exp(\beta' x_\ell)}{\sum_{\ell \in R_i} \exp(\beta' x_\ell)} - \frac{\left[\sum_{\ell \in R_i} x_{\ell j} \exp(\beta' x_\ell)\right]\left[\sum_{\ell \in R_i} x_{\ell k} \exp(\beta' x_\ell)\right]}{\left[\sum_{\ell \in R_i} \exp(\beta' x_\ell)\right]^2} \right] \quad (5.8)$$

for $j, k \in \{1, 2, \ldots, p\}$. A weighted estimate for β is usually obtained by using a root-finding algorithm, such as the Newton-Raphson method. The solution $\hat{\beta}_w$ of (5.7) does not place equal weight to all periods of time, but if proportionality holds then it should be close to $\hat{\beta}$, the MPLE of the original Cox PH model[5]. Of course, the results of this approach depend mainly on the choice of weighting function. According to (5.7) and (5.8) the weighting function can differ from one covariate to another since a mixture of variables with proportional and non-proportional hazards is typical. The choice for each covariate is made based on a preliminary analysis of the proportionality of hazards. In the literature, various options have been mentioned. Some of them are presented below:

- *Gehan scores:* The size of the risk set R_i at event time t_i is used as a weighting function. It is not considered as a very good option since it can lead to low power (Gehan, 1965; Schemper, 1992).

- *Prentice scores:* If n is the total sample size and $\hat{S}(t)$ is the KM estimate of the survivor function based on the whole dataset, another option is to set $w(t_i) = n \cdot \hat{S}(t_i)$, irrespective of the covariate (Prentice, 1978).

- *Xu & O'Quigley's proposal:* Let $P(t)$ be the probability of still being followed-up at t. Xu & O'Quigley (2000) suggested to use $w(t) = [\hat{P}(t)]^{-1}$ as a weighting function. $\hat{P}(t)$ is estimated implementing the KM method, with inverse meaning of the status indicator δ_i. This approach is considered to yield time-averaged regression effects and it has been repeatedly praised in the literature. Under non-PH, it has been shown that the typical Cox PH model estimate for the HR depends on the censoring distribution (Struthers & Kalbfleisch, 1986; Nguyen & Gillen, 2012), even though without censoring it has the interpretation of a time-averaged effect despite the validity of the PH assumption. However, Xu & O'Quigley's (2000) suggestion gives an estimate which is asymptotically independent of the censoring distribution and at the same time equal to the MPLE of the conventional Cox model under PH.

- *Boyd, Kittelson & Gillen's proposal:* It is quite similar to the previous, since $\hat{\beta}_w$ has the same properties: it is equal to the Cox PH model's estimate under

[5]Recall the proportionality test suggested by Lin (1991), discussed in section 3.2.2.

proportionality, but it is robust against the censoring distribution when PH assumption does not hold (Boyd et al., 2012; Rufibach, 2019). Here, the weighting function is the inverse of the probability of still being followed-up at a certain time point t given some special characteristics, meaning that a different censoring distribution can be assumed for each group of individuals sharing the same covariate values. In a sense, Xu & O'Quigley's (2000) approach is a special case of this method, if one assumes that the censoring distributions of all groups are identical.

- *Schemper, Wakounig & Heinz's proposal:* Schemper et al. (2009) suggested the weighting function $w(t) = \hat{S}(t)[\hat{P}(t)]^{-1}$ which results in an average HR. These authors have shown that in a two-sample comparison, average hazard ratios approximate the odds of concordance very well, i.e.,

$$\text{HR} \approx \text{OC} = \frac{c}{1\text{–}c} = \frac{\text{P}(T_1 < T_2)}{\text{P}(T_2 < T_1)}$$

where T_1 and T_2 are the survival times of two randomly chosen subjects of groups 1 and 2. When $\hat{\beta}_w$ is estimated, c can be computed by

$$c = \frac{\exp(\hat{\beta}_w)}{1 + \exp(\hat{\beta}_w)}.$$

This method is suitable for decision-making since it provides a single measure of relative risk summarizing the nature of the data (Dunkler et al., 2018).

- *FH family:* In section 5.2.2, a wide variety of versatile log-rank tests were presented. Lin et al. (2020) suggested performing a combination of weighted log-rank tests, and if the null hypothesis of identical survival functions is rejected based on the max-combination test, then the weighting function from the FH family corresponding to the individual test with the lowest p-value should be used to implement a weighted Cox regression. In this way, more importance is given to time periods which seem to display the greatest difference regarding the survival profile of the groups of interest.

Inference about the weighted estimates can be based on the corresponding covariance matrix, which can be computed employing several approaches according to the literature (Schemper, 1992; Schemper et al., 2009; Boyd et al., 2012; Dunkler et al., 2018). The most popular amongst them are the following:

- *Lin & Sasieni's sandwich estimate:* According to Lin (1991) and Sasieni (1993) the covariance matrix of the weighted estimates can be computed by

$$V = A^{-1}BA^{-1},$$

where

$$A_{jk} = \sum_{i=1}^{n} w(t_i)\frac{-\partial^2\ell(\beta)}{\partial\beta_j\partial\beta_k}$$

and

$$B_{jk} = \sum_{i=1}^{n}[w(t_i)]^2\frac{-\partial^2\ell(\beta)}{\partial\beta_j\partial\beta_k}$$

for $j,k \in \{1,2,\ldots,p\}$. When $w(t) = 1$ this estimate reduces to the inverse of the Fisher information matrix of the Cox PH model. Unfortunately, it is valid only under proportionality of hazards and without model misspecification.

- *Therneau & Grambsch's alternative approach:* An alternative definition of the covariance matrix was given by Therneau & Grambsch (2000), according to which

$$V = A^{-1}(U'U)A^{-1}.$$

The (i,j)-th element of U is

$$\begin{aligned} U_{ij} &= (1-\delta_i)w(t_i)\left[x_{ij} - \frac{\sum_{\ell\in R_i} x_{\ell j}\exp(\beta' x_\ell)}{\sum_{\ell\in R_i}\exp(\beta' x_\ell)}\right] \\ &- \sum_{i':t_{i'}\le t_i}(1-\delta_{i'})w(t_{i'})\frac{\exp(\beta' x_i)}{\sum_{\ell\in R_{i'}}\exp(\beta' x_\ell)}\left[x_{ij} - \frac{\sum_{\ell\in R_{i'}} x_{\ell j}\exp(\beta' x_\ell)}{\sum_{\ell\in R_{i'}}\exp(\beta' x_\ell)}\right] \end{aligned}$$

for $i \in \{1,2,\ldots,n\}$ and $j \in \{1,2,\ldots,p\}$. This estimate is identical to the sandwich variance estimator proposed by Lin & Wei (1989), which they have shown to be robust against non-PH.

- *Jackknife method:* Finally, a variance estimator can occur by estimating the regression coefficient leaving out each individual in turn. Let J be a $n \times p$ matrix with i-th row equal to

$$J_i = \hat{\beta}_w - \hat{\beta}_w^{(i)}$$

where $\hat{\beta}_w^{(i)}$ is the solution of (5.7) if the i-th individual is not included in the model. Then an estimator for the covariance matrix is

$$V = \frac{n-1}{n}(J-\bar{J})'(J-\bar{J}),$$

where $\bar{J}$ is the matrix of column means of J.

A comparison of these three methods, led to conclusion that the robust estimate (second approach) seems to perform better than the other two: the Jackknife estimator has the smallest bias but it requires significantly more time to be calculated than the other two, while at the same time, Lin and Sasieni's method is considered invalid under non-PH. Thus, a good compromise between bias and efficiency, according to Schemper et al. (2009), is the second approach.

Weighted Cox regression is easily implemented in `R`: the package coxphw calculates the weighted estimates proposed by Xu & O'Quigley (2000) and Schemper et al. (2009). Also, inference about the significance of the treatment effect for the two-sample case can be made via the robust variance estimator or the Jackknife method.

5.3.4 Cauchy combination of change-point Cox regressions

In a recent paper by Zhang, Li, Mehrotra & Shen (2021) an innovative omnibus test for the significance of treatment effect has been proposed, using a combination of multiple single change-point Cox regression models. A simulation study where various non-PH patterns were considered showed that this particular approach has robust power against various types of departure from proportionality and at the same time, it controls the type I error at very stringent levels of significance, such as 10^{-4}. Apart from that, it is an easily implemented and comprehensible method, which has the ability to provide a suitable change-point Cox model for the data, if the null hypothesis of no treatment effect is eventually rejected.

A single change-point Cox model for the two-sample case problem, is given by

$$\lambda_i(t) = \lambda_0(t) \exp[\beta(t)' x_i] \tag{5.9}$$

where

$$\beta(t) = \begin{cases} \beta_1, & \text{if } 0 < t < t_{CP} \\ \beta_2, & \text{if } t \geq t_{CP}. \end{cases}$$

Therefore, the null hypothesis to be tested is

$$H_0 : \beta_1 = \beta_2 = 0.$$

In order to fully understand the CauchyCP testing procedure, its steps are presented bellow, one by one:

1. To begin with, a set of m candidate change points $t_1, t_2, \ldots, t_m$ is selected. Usually, $t_1 = 0$ so as to include the Cox PH model in the multiple testing procedure.

2. For each change point t_j a single change-point model such as the one in (5.9) is fitted, with $t_{CP} = t_j, j = 1, 2, \ldots, m$.

3. A likelihood ratio test is conducted to test the null hypothesis $H_0 : \beta_{j1} = \beta_{j2} = 0$ for each $j \in \{1, 2, \ldots, m\}$, separately. The corresponding p-value is denoted by p_j.

4. The individual p-values are combined in a single test with final p-value

$$p_{\mathrm{c}} = 0.5 - \frac{\tan^{-1}(\mathrm{c})}{\pi},$$

where $\mathrm{c} = \sum_{j=1}^{m} \tan[\pi(0.5 - p_j)]/m$. The combination statistic has an asymptotic standard Cauchy distribution regardless of the correlation of the individual p-values.

If there is some a-priori knowledge about the non-PH pattern of the data the sequence of change points must be determined accordingly. However, this is rarely the case, and for that reason it is oftentimes suggested to choose time points covering the whole range of the event times. For instance, one can choose four candidate change-points $t_1 = 0$ and t_2, t_3, t_4 as the 25-th, 50-th and 75-th percentiles of the event times, respectively. The idea behind the proposed CauchyCP method is that, although the majority of the candidate change points are likely misspecified, at least one of them is close to the true value. Thus, by combining the p-values of these change-point models, the treatment effect under non-proportional hazards can be adequately detected with properly controlled type I error. If the null hypothesis is rejected, then the time point corresponding to the smallest individual p-value is chosen and a change-point Cox model is fitted to the data, providing two distinct HR estimates, one representing the time period up to the selected t_{CP}, and another for the subsequent time interval.

5.4 Restricted Mean Survival Time

5.4.1 Definition and properties

The usage of weighted parameter estimations for reporting a single summary measure in cases of non-proportional hazards has provoked controversy in the statistical community, with Royston & Parmar (2011) being the main disputants. As an alternative measure of overall treatment effect, they suggested the Restricted Mean

Survival Time (RMST), a quantity initially introduced by Irwin (1949) but overlooked for years. Essentially, RMST is the mean of survival time up to a fixed time cut-point τ and can be interpreted as "τ-year life expectancy". It is inseparably connected with the survival function as it is equal to the area under the survival curve. Indeed, if $R = \min(T, \tau)$, where T is a random variable denoting the survival time of an individual and τ is a specified time point of interest, then the RMST is defined as follows:

$$\text{RMST}(\tau) = E[R] = E[\min(T, \tau)]. \tag{5.10}$$

R is a non-negative random variable taking values ranging from 0 to τ. Therefore, its mean can be computed by the formula

$$E[R] = \int_0^\tau [1 - F_R(u)]du \tag{5.11}$$

where

$$F_R(u) = P(R \leq u) = P(\min(T, \tau) \leq u) = P(T \leq u) = F_T(u),$$

for $u \in [0, \tau]$. Consequently, (5.11) becomes

$$E[R] = \int_0^\tau [1 - F_R(u)]du = \int_0^\tau [1 - F_T(u)]du = \int_0^\tau S_T(u)du$$

and thus, according to (5.10),

$$\text{RMST}(\tau) = \int_0^\tau S_T(u)du. \tag{5.12}$$

In the literature, one can identify three basic properties of the RMST:

1. It is an increasing function of the chosen time point τ,
2. the limit of the RMST(τ) as $\tau \to \infty$ is equal to the unrestricted mean survival time, which is difficult and in many cases impossible to estimate due to censoring, and as a consequence,
3. the RMST is always smaller than the mean survival time.

Finally, it should be mentioned that instead of the RMST another quantity can be used to express the survival profile of a population: the Restricted Mean Time Lost (RMTL). This quantity is defined as the expected value of $\tau - R$, i.e.,

$$\text{RMTL}(\tau) = E[\tau - R] = \tau - E[\min(T, \tau)] = \int_0^\tau [1 - S_T(u)]du.$$

Of course, if the RMST(τ) is known then the RMTL(τ) can be calculated directly and vice versa. Since the RMST up to a time point τ has a slightly easier interpretation than the corresponding RMTL, it is the one most usually reported in papers.

5.4.2 Estimation from the data

In practice, a non-parametric estimate for RMST(τ) can be obtained by combining (5.12) and the KM estimator of the survivor function. For simplicity, RMST(τ) will be denoted by $\varphi(\tau)$. If there are m failure times before τ, then

$$\hat{\varphi}(\tau) = \sum_{j=1}^{m} \hat{S}(t_{j-1})[t_j - t_{j-1}] + \hat{S}(t_m)[\tau - t_m],$$

where $t_0 = 0$. This estimator is unbiased and its standard error is equal to

$$\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^{m} d_j}{\sum_{j=1}^{m} d_j - 1} \sum_{j=1}^{m} \frac{d_j A_j^2}{r_j(r_j - d_j)}},$$

where d_j is the number of events at t_j, r_j is the number of subjects at risk at t_j and $A_j = \int_{t_j}^{\tau} \hat{S}(t)dt$. When two competing treatments are to be compared, the difference (or the ratio) between the RMSTs of the randomized arms can be used and a test statistic can be calculated. Generally, for K groups, the null hypothesis of no difference can be expressed as

$$H_0 : \varphi_1(\tau) = \varphi_2(\tau) = \ldots = \varphi_K(\tau),$$

while the alternative is

$$H_1 : \varphi_{i_1}(\tau) \neq \varphi_{i_2}(\tau),$$

for some $i_1, i_2 \in \{1, 2, \ldots, K\}$. Let Σ be the covariance matrix of the vector $\varphi(\tau) = (\varphi_1(\tau), \varphi_2(\tau), \ldots, \varphi_K(\tau))'$. Then Σ is a diagonal matrix with diagonal elements the quantities $\hat{\sigma}_j^2$. Let also D be a $(K-1) \times K$ matrix whose j-th row is $e_j - e_{j+1}$, where e_j is a K-dimensional vector whose j-th element is equal to 1 and all others are equal to zero. Then, the test statistic is

$$\varphi(\tau)'[D'(D\Sigma D')^{-}D]\varphi(\tau)$$

and it asymptotically follows a chi-square distribution with rank($D\Sigma D'$) degrees of freedom. This homogeneity test does not identify which pairs are different and thus, if H_0 is rejected, pairwise comparisons are being performed, adjusting the p-values to avoid falsely significant results. For instance, the statistical software `SAS` uses a well-known method, called Šidák's (1967) correction.

Of course, apart from the aforementioned non-parametric method for the calculation and the comparison of the RMST amongst different groups of interest, various parametric models have also been developed. The simplest formulation one

can think of, is the linear model with response the RMST and the individual characteristics as exploratory variables. However, RMST is non-negative and thus a linear model might yield estimates out of bounds. It is usually preferable to fit a log-linear model so as to avoid having an uninterpretable estimate for the RMST. Modeling RMST via a parametric model permits the adjustment for many covariates simultaneously. If the linear model is implemented then the effects are interpreted as differences in the RMST, while if the log-linear model is fitted to the data, then the effects are interpreted in terms of RMST ratios. Due to the nature of the data in survival analysis, i.e., due to the fact that some of the observations are censored, modeling of RMST is accomplished using either pseudo values or Inverse Probability Censoring Weighting (IPCW), with the first method assuming that censoring is not informative, and the second that the censoring distribution can be properly estimated. Nevertheless, other approaches have also been proposed, based on the fact that RMST can be easily computed if the survival function of interest is estimated (see for example Royston & Parmar's (2002) method based on their flexible hazard scaled family of models).

Under the two-sample problem, whether the approach used is non-parametric or parametric, the estimate of the RMST and its standard error provide important information about the significance of the treatment effect and an appropriate test can be performed in a conventional manner.

5.4.3 Choice of τ

It is evident that the results obtained by any model for the RMST are dependent on the choice of the time point τ. Usually, the selected τ is close to the end of the follow-up. For instance, when there are two populations in the study, τ may be set equal to the minimum of the largest observed event time in each of the two treatment groups, or equal to the minimum of the largest observed event or censoring time. Other approaches, including the choice of a time point τ which has some clinical relevance or a trial-specific τ, have been developed and presented in recent papers focusing on the design stage of the study. In general, one should keep in mind that τ should be selected according to the problem at hand and the accumulated information, otherwise invalid findings, such as biased or unstable estimates, will occur.

5.4.4 Combined test by Royston & Parmar

The RMST method is increasingly being considered as an alternative analysis approach when non-PH are apparent. However, due to the fact that the Cox PH model exhibits maximum power under proportionality, Royston & Parmar (2016) suggested a combination of the two methods to gain an improved statistic for the treatment effect testing. They acknowledged the fact that the choice of τ plays a major role on the outcome of the RMST test, and that is why they proposed implementing the corresponding test on a range of τ values and use the maximum statistic instead of the traditional square of the ratio of the RMST difference at a pre-specified time point to its standard error. A suitable adjustment of the resulting minimal p-value is accomplished via a permutation test and finally, it is combined with Cox PH model's p-value using a multiple testing correction closely related to the one introduced by Bonferroni. More specifically, the algorithm steps in order to obtain the final statistic are described as follows:

1. Firstly, a grid of time values for the calculation of the RMST difference statistic must be selected: Since it is unlikely to obtain a reliable representative and clinically meaningful estimate of the RMST difference early in follow-up, the lower bound should not be too small. In the relative paper, the 30th centile of the event times is considered as a reasonable choice for the lower bound (τ_{start}). For the upper bound, a logical choice is the minimum of the largest uncensored event times in the two arms (τ_{end}). As for the number of time points on which the RMST difference statistic is calculated, it is somewhat arbitrary. Royston & Parmar have shown, based on twenty non-randomly chosen trial datasets, that 5 points usually miss the optimal τ, but 10 seem to be enough, since the performance was not quite different from when more points were selected. So, 10 equally spaced times are being selected, i.e.,

$$\tau_k = \tau_1 + \frac{\tau_{10} - \tau_1}{9} \cdot (k-1)$$

   for $k = 1, 2, \ldots, 10$, where $\tau_1 = \tau_{\text{start}}$ and $\tau_{10} = \tau_{\text{end}}$.

2. The RMST difference statistic is calculated for each $\tau_k, k = 1, 2, \cdots, 10$. The maximum value amongst them is denoted by $C_{\max}$ and the corresponding p-value $p_{\max}$ obtained from a chi-square distribution with 1 d.f. is the minimum.

3. A permutation test is implemented in order to gain a corrected version of the previous p-value, since multiple tests have been performed. Firstly, the

treatment covariate is permuted M times in order to remove any systematic association between the treatment assignment and the outcome, preserving the structure of the data. In each permuted dataset step 2 is applied, i.e. the maximal chi-square statistic C_i over 10 selected and equally spaced times is calculated, resulting in a sample of M values from the null distribution of $C_{\max}$. Then, a corrected p-value occurs as follows:

$$p_{\text{perm}} = \frac{N + 0.5}{M + 1},$$

where $N = \sum_{i=1}^{M} I(C_i > C_{\max})$ and 0.5 is a continuity correction. The smallest p_{perm} is equal to $0.5/(M + 1)$. The definition of p_{perm} is quite reasonable: the smaller the N, the smaller the p-value, because then it is rarer to find a chi-square statistic which is greater than $C_{\max}$. Usually, M is set equal to 999 in simulation studies, but in definitive analysis it should be larger.

The aforementioned method exhibits three main disadvantages: it is time-consuming, stochastic and thus not precisely reproducible and finally, the choice of M is arbitrary. An approximation of the p_{perm} can occur using its relationship with $p_{\max}$ given by Royston & Parmar (2016), which is based on a Bod-Tidwell model of the form $E(y) = \beta_1 x^{p_1} + \beta_2 x^{p_2}$. Employing three example datasets, it was shown that

$$E(p_{\text{perm}}) = 1.762(p_{\max})^{0.885} – 0.802(p_{\max})^{2.547}. \tag{5.13}$$

In general, after checking the validity and accuracy of this approximation, they came to the conclusion that it performs quite well. They proposed, however, to implement the accurate method when the approximation in (5.13) seems to be very close to critical values, such as 0.05.

4. After the computation of p_{perm}, the RMST and the Cox test must be combined. Nevertheless, they are positively correlated since both tests correspond to departures from the null hypothesis of identical survival functions. As a result, the min value $p_{\min}$ of the corresponding p-values p_{perm} and p_{Cox} will be significant too often. In this case, a correction for multiple testing procedures should be applied. Here, another empirical approach is used to approximate the null distribution of $p_{\min}$ based on the idea that it is a two parameter beta distribution, to allow some flexibility. Notice, that another important reason for this choice is the set of possible values for $p_{\min}$. The two parameters are

estimated via the maximum likelihood. Eventually, the final p-value denoted by p_{comb} is given by the formula[6]:

$$p_{\text{comb}} = F(p_{\min})$$

where F is the cumulative distribution function of a Beta random variable with parameters $a = 1$ and $b = 1.5$.

When the null hypothesis is rejected, we can suspect the reason by examining the p-values p_{perm} and p_{Cox}. The smallest will show the dominant problem, but it is also useful to do some extra analysis such as the GT test for proportionality and/or the smoothed scatter plots of the scaled Schoenfeld residuals.

Various modifications of the combined test can be considered, mostly replacing the Cox (1972) test with a weighted log-rank test (see section 5.2). In this way, the (weighted) combination places more importance on a specified time period, in order to detect early, late or middle difference between the survival curves of the two arms more easily. The procedure is exactly the same, except from the last step, where p_{Cox} must be replaced with the p-value p_{WLT} from a weighted log-rank test. The parameters of the beta distribution above should be modified accordingly.

5.5 Weighted Kaplan-Meier Statistics

A natural way to perform a test for treatment effect is to directly compare the survivor function estimates of the two populations of interest. Pepe & Fleming (1989, 1991) presented a class of tests called *Weighted Kaplan-Meier tests*. The initial idea was to conduct a test for treatment effect based on the quantity

$$T(\tau) = \int_0^\tau [\hat{S}_1(t) - \hat{S}_2(t)]dt, \tag{5.14}$$

where τ is the length of the study period. However, in the presence of heavy censoring the difference between the survivor curves can be very unstable for t close τ. Notice that, according to (5.12),

$$T(\tau) = \int_0^\tau [\hat{S}_1(t) - \hat{S}_2(t)]dt = \text{RMST}_1(\tau) - \text{RMST}_2(\tau)$$

and thus, this test is equivalent to the test presented in section 5.4.2. This is the reason why poor choices of τ result in low power.

[6]Both approximations used in the combined test were shown to be adequate for practical application.

To overcome this problem, Pepe & Fleming (1989, 1991) proposed to base the test on the interval of a weighted difference of the survivor functions, i.e., on the quantity

$$T_w(\tau) = \int_0^\tau w(t)[\hat{S}_1(t) - \hat{S}_2(t)]dt. \tag{5.15}$$

Various weighting functions $w(t)$ can be used but the aim here is to choose one that ensures the stability of the statistic. A famous suggestion, is to use the harmonic mean of the probabilities $C_j(t)$, $j = 1, 2$, of no censoring before time t for the two groups. More precisely,

$$w_C(t) = \frac{C_1(t)C_2(t)}{p_1C_1(t) + p_2C_2(t)},$$

where p_1 is the proportion of patients in sample 1 and p_2 is the proportion of patients in sample 2. In the absence of censoring, $w_C(t) = 1$.

Similar to the RMST difference, the quantity $T_w(\tau)$ devided by its standard deviation $\sigma_{T_w}(\tau)$ has an asymptotic standard normal distribution. It holds that the variance of $T_w(\tau)$ is given by

$$\sigma^2_{T_w}(\tau) = -\frac{n}{n_1n_2}\int_0^\tau \frac{[\int_t^\tau w(u)S(u)du]^2}{S^2(t)}[w_C(t)]^{-1}dS(t),$$

where n is the total number of observations and n_j is the number of patients in group j, $j = 1, 2$.

Chapter 6

Simulation study: Tests for treatment effect

6.1 Data simulation: Special scenarios

The aim of this chapter is to compare the performance of various tests for treatment effect presented in Chapter 5. Therefore, for two populations with survivor functions $S_1(t)$ and $S_2(t)$, the null hypothesis H_0 is expressed as $S_1(t) = S_2(t)$, while the alternative H_A as $S_1(t) \neq S_2(t)$. For the null hypothesis of identical survivor functions, two sample sizes for the total number of patients are being under consideration: $n = 200$ and $n = 1000$. In each case, the patients are distributed equally between the control and the intervention group and their survival time follows an exponential distribution with rate $\lambda = 1$. For the alternative hypothesis of dissimilar survivor profiles between the subjects of the two arms, all the scenarios discussed in Chapter 4 are investigated. The proportion of randomly censored observations reaches 5% in the whole data set and the number of repetitions is equal to 1000 for each scenario.

6.2 Results

Twenty tests for treatment effect are being compared in this chapter:

1. a max combination of weighted log-rank tests, using the sequence $0, 0.1, 0.2, ..., 0.9, 1$ for ρ and γ ($11 \times 11 = 121$ individual z-statistics),

2. the max combination test by Lin et al. (2020),

3. Karrison's (2016) versatile weighted log-rank test,

4. Lee's (1996) versatile weighted log-rank test,
5. Lee's (2007) versatile weighted log-rank test,
6. a weighted log-rank test from the FH family (see section 5.2.1), with $\rho = 1$ and $\gamma = 0$ (Log-rank for early effects - LRE),
7. a weighted log-rank test from the FH family (see section 5.2.1), with $\rho = 0$ and $\gamma = 1$ (Log-rank for late effects - LRL),
8. the traditional log-rank test (see sections 2.3 and 5.2.1),
9. Cox's (1972) test for the significance of the treatment indicator variable,
10. joint test by Royston & Parmar (2014),
11. Breslow, Elder & Berger's (1984) combination test using rank scores,
12. the supremum log-rank test, which is essentially a combination of the traditional log-rank test with itself, since the log-rank statistic is calculated up to each failure time and the maximum of all these statistics is set to be the definitive statistic for the final test (Fleming et al., 1987),
13. the RMST difference using the minimum of the maximum observed failure or censoring times in the two arm (see section 5.4),
14. the RMST difference using the minimum of the maximum observed failure times in the two arms (see section 5.4),
15. the Combined test by Royston & Parmar (2016),
16. a weighted version of the aforementioned test, using the LRL instead of Cox's test,
17. the Weighted KM test (Pepe & Fleming, 1989, 1991),
18. the Cauchy CP testing procedure (Zhang et al., 2021),
19. Weighted Cox Regression employing the weights proposed by Schemper et al. (2009), resulting in an average hazard ratio (AHR), and finally,
20. Weighted Cox Regression using the weights proposed by Xu & O'Quigley (2000), giving an average regression effect (ARE).

The results per each scenario are presented below, while additional tables and figures for further insight on the findings are included in the Appendix Section B.

Identical Survivor Functions

The empirical significance level under the null hypothesis of identical survivor functions is, as expected, approximately equal to 5% (Table 6.1, Figure 6.1). As the sample size increases from 200 to 1000, type I error decreases for the majority of the tests.

Test	$n = 200$	$n = 1000$	Test	$n = 200$	$n = 1000$
1	5.2	4.5	11	5.5	4.8
2	5.3	4.4	12	6.1	5.8
3	6.2	4.8	13	5.7	5.7
4	6.0	4.8	14	4.9	5.2
5	6.0	4.6	15	5.8	4.8
6	6.4	5.1	16	6.4	5.6
7	5.0	5.6	17	5.7	5.7
8	6.1	5.8	18	5.9	5.6
9	6.0	5.8	19	6.4	5.0
10	5.5	4.8	20	6.0	5.9

Table 6.1: Type I error (size in %) of 20 tests for treatment effect, for two different sample sizes n.

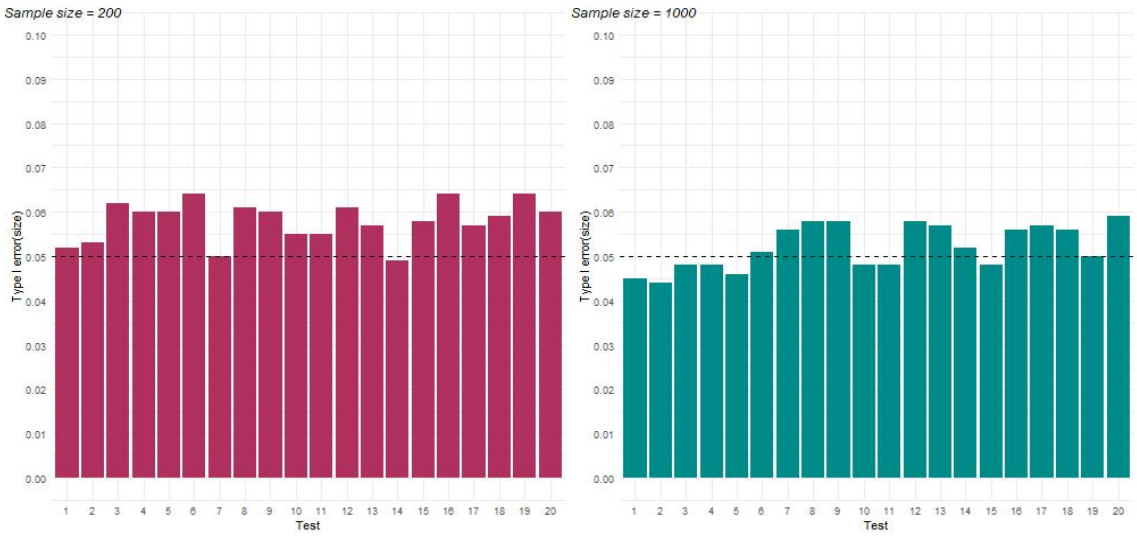

Figure 6.1: Type I error (size) of 20 tests for treatment effect, for each sample size n. The dashed line corresponds to type I error equal to 5%.

	Hazard Ratio					
	0.65		0.8		0.9	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	82.4	100	29.9	90.3	10.7	34.4
2	82.6	100	30.1	90.5	11.0	34.4
3	82.7	100	30.3	90.3	11.3	34.4
4	79.6	100	28.1	88.4	10.5	32.2
5	80.8	100	29.6	89.7	11.0	33.9
6	72.8	100	25.7	83.0	10.7	30.1
7	75.4	100	27.8	83.7	9.3	29.6
8	85.2	100	31.1	92.7	10.7	35.7
9	85.2	100	30.9	92.7	10.5	35.7
10	77.3	100	25.3	87.0	9.7	29.0
11	77.2	100	25.4	87.0	9.8	28.9
12	84.8	100	31.0	92.5	10.3	35.5
13	85.6	100	31.1	92.8	10.7	35.6
14	83.4	100	28.1	92.3	9.3	35.4
15	81.8	100	28.4	90.4	10.5	32.2
16	80.8	100	30.7	89.3	11.2	34.2
17	85.6	100	31.1	92.6	10.8	35.8
18	81.3	100	28.1	89.7	10.8	32.6
19	73.3	100	26.2	83.9	10.9	30.4
20	85.3	100	30.8	92.8	10.9	35.8

Table 6.2: Power(%) of 20 tests for treatment effect under the proportional hazards assumption, using three constant HR functions and two different sample sizes n.

Proportional Hazards

Under the assumption of proportional hazards, as the assumed HR decreases so does the power of the tests. Of course, the log-rank test along with Cox's test for significance[1] achieve the maximum possible power, although in same cases they are slightly outperformed by Tests 12, 13, 17 or 20. This is only due to the fact that the data are simulated only 1000 times. This means, however, that the supremum log-rank test, the RMST difference using as τ the minimum of the maximum observed times, the weighted KM test and the weighted cox regression resulting in an ARE have comparable power with the traditional log-rank test under PH. On the other hand, LRE, LRL, joint test by Royston & Parmar (2014) and Breslow's (1984) test, along with the AHR parameters of the weighted cox regression show the greatest lack of power in comparison to the other methods (Table 6.2, Figure 6.2).

[1]These tests are considered equivalent (see section 2.4.4).

Figure 6.2: Power of 20 tests for treatment effect under the PH assumption, for each sample size and HR.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	9.4	31.9	14.5	57.5	21.0	76.3
2	9.9	32.5	15.0	57.2	20.3	75.7
3	10.3	33.6	15.1	58.1	20.7	73.8
4	10.2	34.3	14.4	55.2	19.1	73.1
5	10.7	34.3	14.8	58.0	20.2	74.0
6	12.2	40.8	19.0	64.9	23.2	79.0
7	5.9	8.3	7.3	19.3	12.8	41.5
8	8.6	26.4	14.6	51.1	21.2	74.2
9	8.6	26.2	14.6	51.0	21.2	74.1
10	11.1	36.2	14.9	57.4	18.4	71.6
11	11.1	36.2	15.1	57.3	18.3	71.6
12	8.8	26.4	14.3	51.1	21.2	73.9
13	8.0	24.8	13.8	51.1	21.5	74.7
14	7.7	24.8	13.7	50.3	20.2	74.4
15	11.6	36.9	16.2	58.2	21.4	75.4
16	12.1	34.7	16.2	55.8	20.4	72.6
17	8.7	26.0	14.3	52.2	21.5	75.6
18	10.6	34.5	15.3	57.4	19.7	74.8
19	12.0	39.9	18.9	64.7	23.1	78.9
20	8.4	24.2	13.6	49.1	21.2	72.0

Table 6.3: Power(%) of 20 tests for treatment effect, for the early effect case with initial HR = 0.8 and subsequent HR $\approx$ 1 after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

Early/Diminishing Effect

When an early effect is anticipated, for instance, when the HR is initially equal to 0.8 and subsequently equal to 1, approximately, LRE achieves maximum power (Table 6.3, Figure 6.3). Interestingly, Tests 1 to 5 and 19 also perform quite well in all cases (see Tables B.1 & B.2 and Figures B.1 and B.2 in the Appendix). At the same time, when the change in the HR is at the beginning or the middle of the study, the Combined test by Royston & Parmar (2016), the joint test (Royston & Parmar, 2014) and Breslow's (1984) combination have also good power in comparison to the rest of the tests. However, when the change happens after the occurrence of 70% of the events in the treatment group, the joint and Breslow's tests along with LRL show a severe lack of power. Of course, this holds for all individual cases for the LRL test, since it places more weight on the end of the study, where HR is roughly 1.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	18.7	71.7	12.9	47.2	7.6	23.7
2	19.1	71.9	13.1	47.7	7.7	24.7
3	19.7	71.4	13.6	49.4	8.4	26.7
4	19.0	70.6	13.8	50.1	8.9	27.2
5	18.9	71.1	14.0	49.8	8.5	26.4
6	11.2	33.1	7.4	13.6	7.3	6.8
7	22.8	77.6	16.6	58.9	9.5	31.6
8	16.9	64.8	11.5	38.8	7.3	18.9
9	16.9	64.7	11.2	38.7	7.3	18.9
10	18.5	67.8	13.8	49.7	9.6	28.5
11	18.3	67.8	13.9	49.9	9.4	28.7
12	16.5	64.4	10.5	38.9	7.2	18.7
13	18.2	67.4	11.4	43.5	7.5	21.3
14	15.0	65.4	8.6	40.3	5.6	19.3
15	14.6	59.7	9.7	33.2	7.3	16.5
16	21.1	73.6	15.2	53.1	10.0	27.8
17	17.5	65.5	10.9	41.2	7.7	20.0
18	18.0	69.2	13.5	51.7	9.3	28.8
19	11.3	33.5	7.4	14.5	7.4	7.4
20	17.6	65.7	11.9	41.3	7.2	19.7

Table 6.4: Power(%) of 20 tests for treatment effect, for the late effect case with initial HR ≈ 1 and subsequent HR $= 0.8$ after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

Late/Delayed Effect

Table 6.4 and Figure 6.4 give information about the power of the 20 tests under a late effect scenario, where the final HR is equal to 0.8. In this case, the LRE and the weighted Cox regression using weights from Schemper et al. (2009) have the worst performance. As anticipated, the LRL test and the weighted Combined test by Royston & Parmar (2016) achieve high power in comparison to the others. Cauchy CP testing procedure and max combination tests 1 to 5 exhibit moderate power, but they usually come immediately after LRL and weighted Combined test, with an approximate loss of power of about 3% for small samples and 5-10% for large samples. Similar conclusion are drawn for the cases where the final HR is either equal to 0.65 or 0.9 (see Tables B.3 & B.4 and Figures B.3 & B.4).

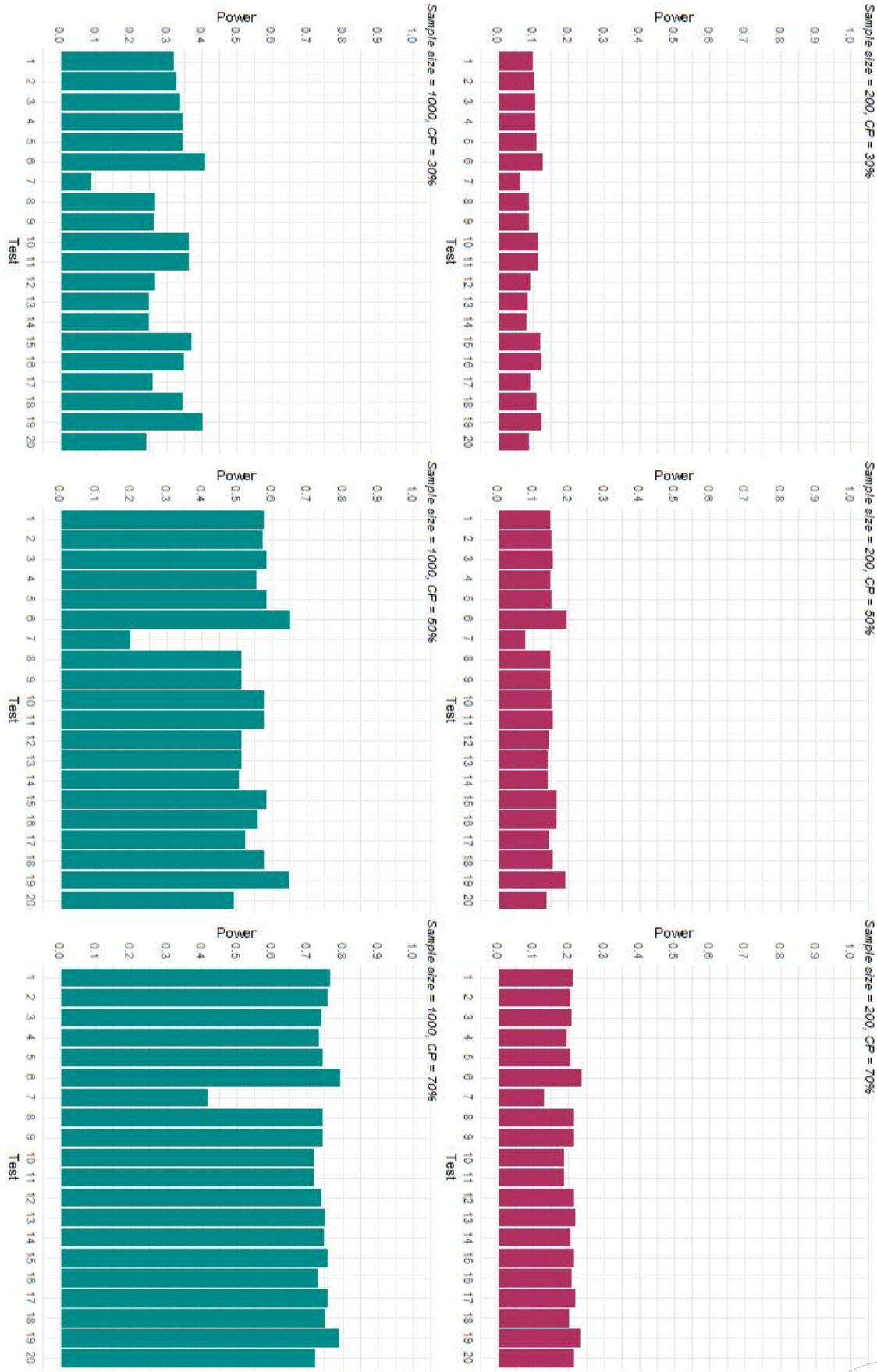

Figure 6.3: Power of 20 tests for treatment effect, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when an early effect with initial HR = 0.8 is observed.

Figure 6.4: Power of 20 tests for treatment effect, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when a late effect with final HR = 0.8 is observed.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	10.1	30.8	10.2	34.3
2	9.7	31.4	10.1	35.3
3	10.2	33.0	10.8	38.1
4	12.0	44.6	12.4	48.9
5	10.8	33.9	11.2	39.4
6	11.4	33.2	10.9	32.2
7	6.1	12.8	8.5	21.1
8	6.3	10.2	6.1	6.1
9	6.2	10.2	6.0	6.1
10	16.5	63.6	17.8	66.5
11	16.5	63.7	17.7	66.4
12	6.2	10.3	6.0	6.2
13	9.2	21.4	6.9	8.2
14	9.3	21.9	5.6	8.1
15	12.2	41.2	12.0	41.1
16	14.9	48.4	14.6	53.0
17	9.2	21.8	6.6	8.0
18	14.6	59.7	14.7	64.9
19	10.7	31.0	10.4	30.2
20	6.0	8.9	5.6	5.6

Table 6.5: Power(%) of 20 tests for treatment effect, for the scenario of crossing hazards with initial HR = 0.8 and subsequent HR = 1.2, for different sample sizes n and cut points 2 and 4.

Crossing Hazards

For both cases of crossing hazards involving a HR equal to 0.8 and a HR equal to 1.2, the joint test (Royston & Parmar, 2014) and Breslow's (1984) proposal with rank scores exhibit the best performance. After them, the Cauchy CP testing procedure also seems a quite reasonable option, followed by the weighted Combined test and Lee's (1996) suggestion (Tables 6.5 & 6.6 and Figures 6.5 & 6.6). The latter outperforms the other max combination tests (1, 2, 3 and 5), despite the fact that it had similar behavior with them up to now.

Many conventional tests have severely diminished power, such as Cox's test, the log-rank test and the RMST difference along with a non-traditional one: the test occurring from a weighted Cox regression with an ARE (Test 20). Moreover, these and other tests (e.g. the supremum log-rank and the weighted KM test) seem to perform even worse when the follow-up period is extended by two time units.

Figure 6.5: Power of 20 tests for treatment effect, for different sample sizes and cut points for the scenario of crossing hazards with initial HR = 0.8 and subsequent HR = 1.2.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	7.6	26.4	8.9	33.2
2	7.5	26.8	9.2	33.8
3	7.8	28.5	9.7	36.0
4	9.9	38.8	11.6	47.8
5	8.3	29.9	10.3	37.9
6	9.0	31.3	9.0	28.8
7	5.2	11.0	8.5	23.6
8	5.6	8.5	6.1	5.2
9	5.4	8.5	6.0	5.2
10	14.4	56.4	17.8	66.0
11	14.2	56.4	17.8	66.1
12	5.5	8.6	5.7	5.1
13	7.1	17.2	5.9	5.3
14	7.4	17.9	5.3	5.3
15	9.1	34.8	9.6	34.0
16	10.6	40.7	13.0	49.5
17	7.1	18.4	5.7	5.7
18	11.6	47.1	14.1	59.2
19	9.3	30.6	8.4	27.8
20	5.3	8.0	5.9	5.9

Table 6.6: Power(%) of 20 tests for treatment effect, for the scenario of crossing hazards with initial HR = 1.2 and subsequent HR = 0.8, for different sample sizes n and cut points 2 and 4.

Similar results can be drawn for the other two cases of crossing hazards, leaving out of the equation the previous statement when the initial HR is 1.10 and subsequently becomes equal to 0.65 (Table B.6, Figure B.6). In general, the power of the 20 tests is higher for the latter scenario and the one where HR = 0.65 at the beginning and equal to 1.10 afterwards (see also Table B.5, Figure B.5).

Long-term Survivors

The simulation study in this case is not so informative, but some intriguing findings occur from the case where the initial HR is 0.8. The LRE test and the weighted Cox regression resulting in an AHR have the lowest power. On the other hand, the LRL and the weighted Combined test by Royston & Parmar (2016) achieve better performance than the other 18 testing procedures, since they place more weight at the end of the study, where the effect is greater. However, the majority of the tests perform well, and they are quite close the optimal choices. For instance,

Figure 6.6: Power of 20 tests for treatment effect, for different sample sizes and cut points for the scenario of crossing hazards with initial HR = 1.2 and subsequent HR = 0.8.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	48.7	99.3	61.6	100.0
2	49.0	99.3	62.0	100.0
3	48.3	99.3	61.9	100.0
4	46.2	99.2	58.3	100.0
5	47.7	99.2	60.4	100.0
6	34.8	94.5	37.4	96.2
7	51.0	99.3	65.2	100.0
8	47.1	98.9	60.9	100.0
9	46.6	98.9	60.8	100.0
10	39.8	99.0	53.8	99.9
11	40.0	99.0	53.6	99.9
12	46.6	98.9	59.9	100.0
13	40.2	97.0	60.3	99.9
14	38.6	96.9	57.3	99.9
15	41.1	98.6	54.9	99.9
16	48.1	99.2	62.9	100.0
17	39.7	96.9	59.5	99.8
18	44.5	99.0	59.7	100.0
19	35.1	94.8	38.2	96.8
20	47.0	98.9	62.2	100.0

Table 6.7: Power(%) of 20 tests for treatment effect, for the scenario of long-term survivors with initial HR = 0.8 and subsequent HR = 0.8^2, for different sample sizes n and cut points 2 and 4.

all max combination tests (Tests 1 to 5), the traditional log-rank, the supremum log-rank, the Cauchy CP test and the weighted Cox regression giving an ARE, work pretty well too (Table 6.7, Figure 6.7). Table B.7 and Figure B.7 in the Appendix give also some insight into the other scenario of long-term survivors.

Generally speaking, there is not an optimal procedure for testing the significance of the treatment effect under various non-PH patterns. However, it is obvious that many procedures work well under different scenarios. For instance, max combination tests perform well in all cases apart from crossing hazards. In these scenarios, their power is quite similar. Therefore, there is no gain when a grid of values for ρ and γ is under consideration, i.e., Test 1 is unnecessary complex. Lin et al. (2020) suggested that after a versatile weighted test is conducted, a weighted estimate for the HR can be estimated using as weight the weighting function resulting in the smallest p-value amongst the individual tests. Of course, a weighted Cox model can also be fitted to the data using either weights proposed by Schemper et al. (2009) or by

Figure 6.7: Power of 20 tests for treatment effect, for different sample sizes and cut points for the scenario of long-term survivors with initial HR = 0.8 and subsequent HR = 0.8^2.

Xu & O'Quigley (2000). The first method is preferred for the early effect scenario, while the latter for the diminishing effect and the long-term survivors case. Another suggestion, is the usage of a change point Cox model based on the Cauchy CP testing procedure, which also performs well here.

In the presence of crossing hazards, only three tests yield regularly valid results: the joint test, Breslow's (1984) combination and the Cauchy CP procedure. Since neither of the first two produces an estimate for the HR or the RMST, Cauchy CP method can be implemented both for testing the treatment effect and providing a piecewise constant HR. Unfortunately, it was shown that even RMST based tests were not suitable for this case, so it may be preferable to report a time dependent HR rather than a non-precise summary measure for the whole study.

Chapter 7

Discussion and further research

This dissertation serves as a general overview of - mainly analytical - tests for proportionality and tests for treatment effect. After a brief clarification of fundamental definitions and statistical methods in survival analysis, various testing procedures, developed since the introduction of the Cox proportional hazards model in 1972 up to now, were presented and examined under four non-PH patterns via simulations. The necessity of finding the most powerful test when the proportionality assumption is violated is the result of recent advancements in oncology therapy, and specifically in immunotherapy. Most randomized clinical trials with a time-to-event outcome are designed assuming proportional hazards of the treatment effect. However, due to new, innovative therapies with unique mechanisms of action, several types of non-proportionality patterns usually occur either as a consequence of different treatment effects in subgroups or due to the treatment itself.

The findings of this thesis are summarized in Tables 7.1 and 7.2 using the same numbering for the tests for proportionality and treatment effect as the one in Chapters 4 and 6, respectively. After investigating their performance via simulation studies, it becomes clear that no test surpasses all the others under different alternatives. Amongst the eighteen tests for proportional hazards, four of them display stable behavior under dissimilar types of departure from the null hypothesis: Grambsch & Therneau's suggestion (1994) using as functions of time either the ranks of the failure times or the Kaplan–Meier estimate of the pooled survivor function, a modification of the goodness-of-fit test proposed by Lin (1991) using as weighted parameter estimators the ones introduced by Schemper et al. (2009) and one of the tests proposed by Breslow et al. (1984). The latter, however, is only useful for the two-sample case. As for the comparison of twenty tests for treatment effect, many testing procedures seem to offer proper results: the versatile weighted log-rank tests, the joint and combined tests by Royston & Parmar (2014, 2016), as well as the com-

bination test suggested by Breslow et al. (1984). Nevertheless, a careful study of the tables and figures presented in Chapter 6 shows the superiority and flexibility of a newly developed method known as the Cauchy combination of change-point Cox regressions (Zhang et al., 2021).

Test	Early Effect	Late Effect	Crossing Hazards	Long-term survivors
1	×	×	×	×
2	×	×	×	×
3	×	×	×	
4	×	×	×	×
5	×	×	×	×
6	×	×	×	
7				
8				
9			×	×
10				
11	×	×	×	×
12	×	×	×	×
13	×	×	×	×
14	×	×	×	×
15	×	×	×	×
16				
17	×	×	×	×
18	×	×	×	×

Table 7.1: Tests for proportionality which perform poorly under each scenario.

Despite the large number of testing procedures discussed, numerous other suggestions have been made throughout the years. As for the tests for proportionality already conducted, many improvements can be made. For instance, the performance of the interval-dependent tests was examined only for the case of two non-overlapping time intervals. More change points and/or partitions of the covariate space can be investigated in further studies. Also, a comparison of global tests will be useful, since in real-life applications, a wide range of characteristics, i.e., variables, is reported for each patient. The proportionality assumption may be violated for any covariate, not just the treatment indicator. At the same time, more alternatives should be simulated, in order to also assess the performance of the tests for treatment effect. Throughout this thesis, it was stressed that the early and the late effect scenarios, along with crossing hazards and long-term survivors are just indicative of what may someone encounter during the analysis of non-proportional data. Various time functions for the hazard ratio may be considered, implementing different distribu-

Test	PH	Early Effect	Late Effect	Crossing Hazards	Long-term survivors
1					
2					
3					
4					
5					
6			×	×	×
7		×		×	
8				×	
9				×	
10					
11					
12				×	
13				×	
14				×	
15					
16					
17				×	
18					
19			×	×	×
20				×	

Table 7.2: Tests for treatment effect which perform poorly under each scenario.

tions for the simulation of the patients' survival time in each arm. For instance, the case of non-monotone hazard ratio may be investigated using a bathtub-shaped HR function or the case of a gradually increasing HR simulating survival times from the Weibull distribution.

After all that, real data should be used to validate the results and ensure the validity of the findings. New observations and issues may arise, but an extremely thorough literature review along with a wide range of simulation studies can lead to the development of new methods or the utilization of old ones, in a new, more efficient way.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Appendix A

Simulation study A

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	9.0	14.3	10.2	21.6	9.3	19.4
2	7.8	22.2	7.1	20.8	6.4	11.3
3	7.7	21.0	9.3	30.6	6.4	10.8
4	7.8	13.8	9.2	20.7	8.2	18.7
5	7.9	22.4	7.3	21.3	6.6	11.5
6	7.6	21.0	9.5	31.0	6.3	11.2
7	8.7	25.6	9.3	27.9	6.9	17.2
8	8.9	24.5	9.3	28.0	6.9	17.7
9	8.2	23.0	8.8	26.3	6.5	16.7
10	8.2	23.9	9.1	27.4	7.0	17.8
11	4.9	13.4	6.5	19.2	6.0	16.9
12	7.6	21.0	8.9	30.6	6.3	10.9
13	7.3	20.8	8.6	30.4	6.1	10.9
14	7.1	21.0	8.7	31.2	6.2	11.1
15	7.6	21.0	8.9	30.6	6.3	10.9
16	8.8	25.3	9.3	27.8	6.8	17.3
17	7.7	14.1	9.3	20.6	8.5	18.5
18	7.9	22.5	7.3	21.4	6.4	11.4

Table A.1: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the early effect case with initial HR = 0.8 and subsequent HR ≈ 1 after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

Figure A.1: Power of 18 tests for proportional hazards, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when an early effect with initial HR = 0.8 is observed.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	6.7	6.5	6.7	7.8	7.2	7.3
2	5.9	7.0	5.6	6.9	5.5	5.6
3	6.9	6.7	6.7	9.9	6.2	5.7
4	5.9	6.4	6.5	7.6	6.0	7.0
5	5.8	7.1	5.5	7.0	5.3	5.4
6	5.9	6.4	6.4	9.5	5.9	5.5
7	6.1	7.4	5.9	9.0	5.9	6.8
8	6.0	7.4	6.1	8.8	5.9	6.7
9	5.1	7.1	5.3	8.0	4.9	6.0
10	6.0	7.3	6.3	8.3	5.4	6.9
11	4.5	6.2	4.5	7.3	4.4	7.3
12	6.6	6.8	6.6	9.8	6.1	5.5
13	6.4	6.6	6.4	9.8	5.9	5.5
14	5.9	6.7	6.3	9.8	6.1	5.5
15	6.6	6.8	6.6	9.8	6.1	5.5
16	6.0	7.4	6.0	9.0	6.1	6.6
17	5.6	5.9	6.0	7.3	6.0	6.7
18	5.4	7.2	5.1	7.1	4.9	5.6

Table A.2: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the early effect case with initial HR = 0.9 and subsequent HR ≈ 1 after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

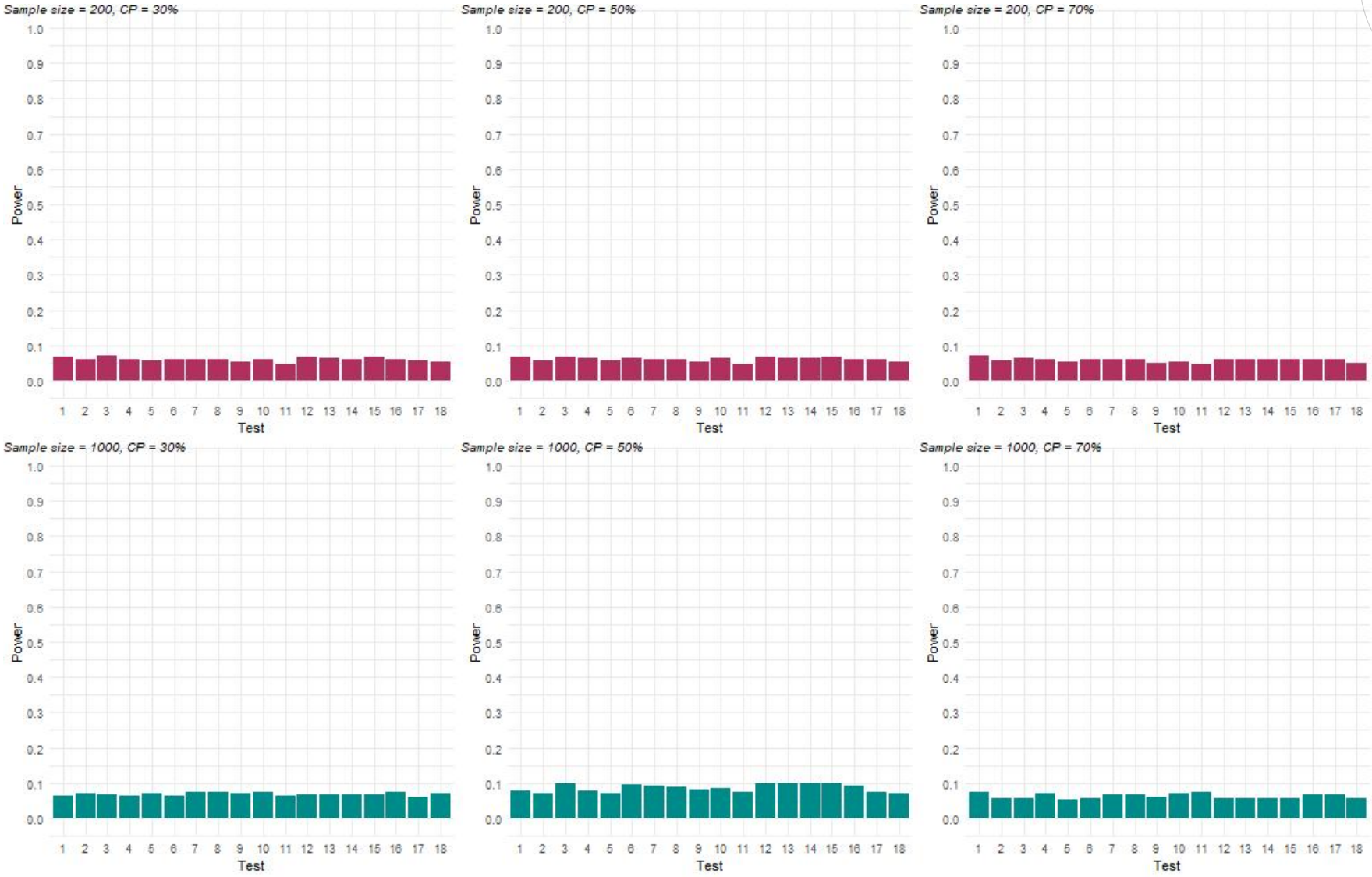

Figure A.2: Power of 18 tests for proportional hazards, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when an early effect with initial HR = 0.9 is observed.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	11.5	17.8	13.8	26.0	13.3	25.7
2	9.4	23.8	9.4	23.9	8.1	16.3
3	9.3	18.8	11.9	36.5	8.0	15.7
4	10.0	17.9	12.1	25.8	11.6	25.6
5	9.3	24.0	9.7	24.4	8.2	16.4
6	9.0	19.3	11.2	35.4	7.3	15.3
7	9.9	24.9	10.9	31.0	8.6	21.1
8	9.9	25.0	11.1	31.1	8.7	21.0
9	9.5	24.1	10.5	29.0	8.2	20.0
10	10.2	25.3	10.9	30.8	8.8	22.3
11	5.2	16.1	6.8	23.6	7.0	23.6
12	9.3	18.4	11.5	36.2	8.0	15.3
13	8.8	18.0	10.7	36.0	7.6	15.2
14	8.6	19.1	11.3	36.8	7.7	16.8
15	9.3	18.4	11.5	36.2	8.0	15.3
16	9.8	24.9	10.9	30.9	8.6	21.0
17	10.2	18.3	12.1	27.0	11.6	25.8
18	9.4	24.4	9.5	24.6	7.7	16.8

Table A.3: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the late effect case with initial HR ≈ 1 and subsequent HR $= 0.8$ after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

Figure A.3: Power of 18 tests for proportional hazards, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when a late effect with final HR = 0.8 is observed.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	8.1	9.6	8.3	11.7	8.8	11.9
2	6.6	8.5	6.4	8.5	6.2	7.0
3	7.5	8.5	7.6	11.9	6.6	7.5
4	6.7	9.3	7.0	11.4	7.8	11.5
5	6.7	8.2	6.5	8.5	5.9	7.1
6	6.4	8.5	6.9	11.8	6.2	7.8
7	5.9	10.1	6.2	10.9	5.8	9.1
8	5.9	10.0	6.2	10.9	5.8	9.3
9	5.9	9.1	6.0	10.3	5.7	8.6
10	5.9	10.8	6.3	11.1	5.9	9.5
11	4.4	9.0	5.0	10.3	5.0	10.2
12	7.3	8.5	7.4	11.7	6.4	7.5
13	7.1	8.5	7.0	11.4	6.1	7.4
14	6.1	8.4	6.8	12.2	5.9	8.5
15	7.3	8.5	7.4	11.7	6.4	7.5
16	5.9	10.1	6.2	10.9	5.7	9.1
17	6.9	9.5	6.9	11.5	7.3	11.2
18	7.0	8.8	6.9	8.8	6.3	6.8

Table A.4: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the late effect case with initial HR ≈ 1 and subsequent HR $= 0.9$ after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

Figure A.4: Power of 18 tests for proportional hazards, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when a late effect with final HR = 0.9 is observed.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	19.5	71.4	18.5	61.8
2	14.7	53.5	16.8	63.1
3	22.6	76.4	25.5	84.3
4	19.6	71.4	18.4	62.1
5	16.4	54.8	17.3	63.9
6	22.6	76.3	25.7	84.3
7	20.6	72.6	21.1	76.3
8	20.2	73.1	21.4	76.2
9	7.7	46.3	20.2	73.7
10	20.2	71.6	21.3	74.4
11	13.0	67.2	13.5	59.6
12	22.4	76.4	25.3	84.4
13	22.4	76.2	24.7	84.3
14	22.7	76.2	24.3	84.1
15	22.4	76.4	25.3	84.4
16	20.2	72.9	21.1	76.3
17	18.8	69.7	17.9	60.8
18	16.6	57.4	18.5	66.2

Table A.5: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the scenario of crossing hazards with initial HR = 0.8 and subsequent HR = 1.2, for different sample sizes n and cut points 2 and 4.

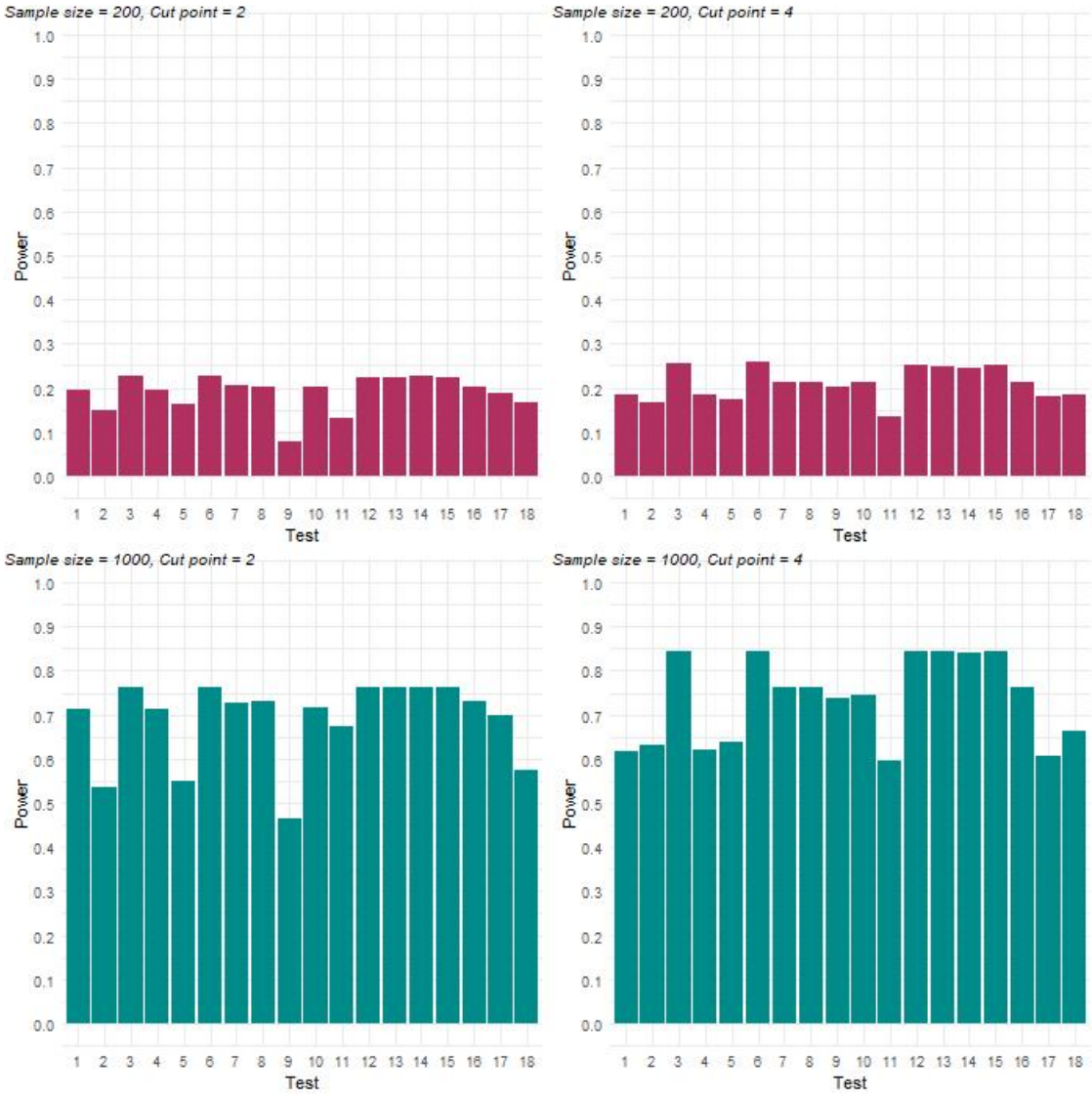

Figure A.5: Power of 18 tests for proportional hazards, for different sample sizes and cut points for the scenario of crossing hazards with initial HR = 0.8 and subsequent HR = 1.2.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	17.8	65.3	21.1	66.6
2	11.8	45.7	16.5	61.7
3	16.3	55.4	21.7	75.3
4	17.8	65.4	21.3	66.8
5	13.3	48.1	17.5	63.2
6	16.2	55.5	21.5	74.8
7	17.1	63.8	22.2	73.1
8	17.3	64.4	22.6	73.1
9	5.6	33.8	20.5	69.4
10	17.4	63.6	23.2	72.8
11	12.9	63.5	14.8	63.5
12	16.3	55.7	21.4	75.1
13	16.1	55.4	21.1	75.1
14	17.1	56.7	21.4	75.7
15	16.3	55.7	21.4	75.1
16	17.1	63.8	22.5	72.8
17	18.4	66.0	21.9	68.0
18	13.6	47.7	17.9	61.8

Table A.6: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the scenario of crossing hazards with initial HR = 1.2 and subsequent HR = 0.8, for different sample sizes n and cut points 2 and 4.

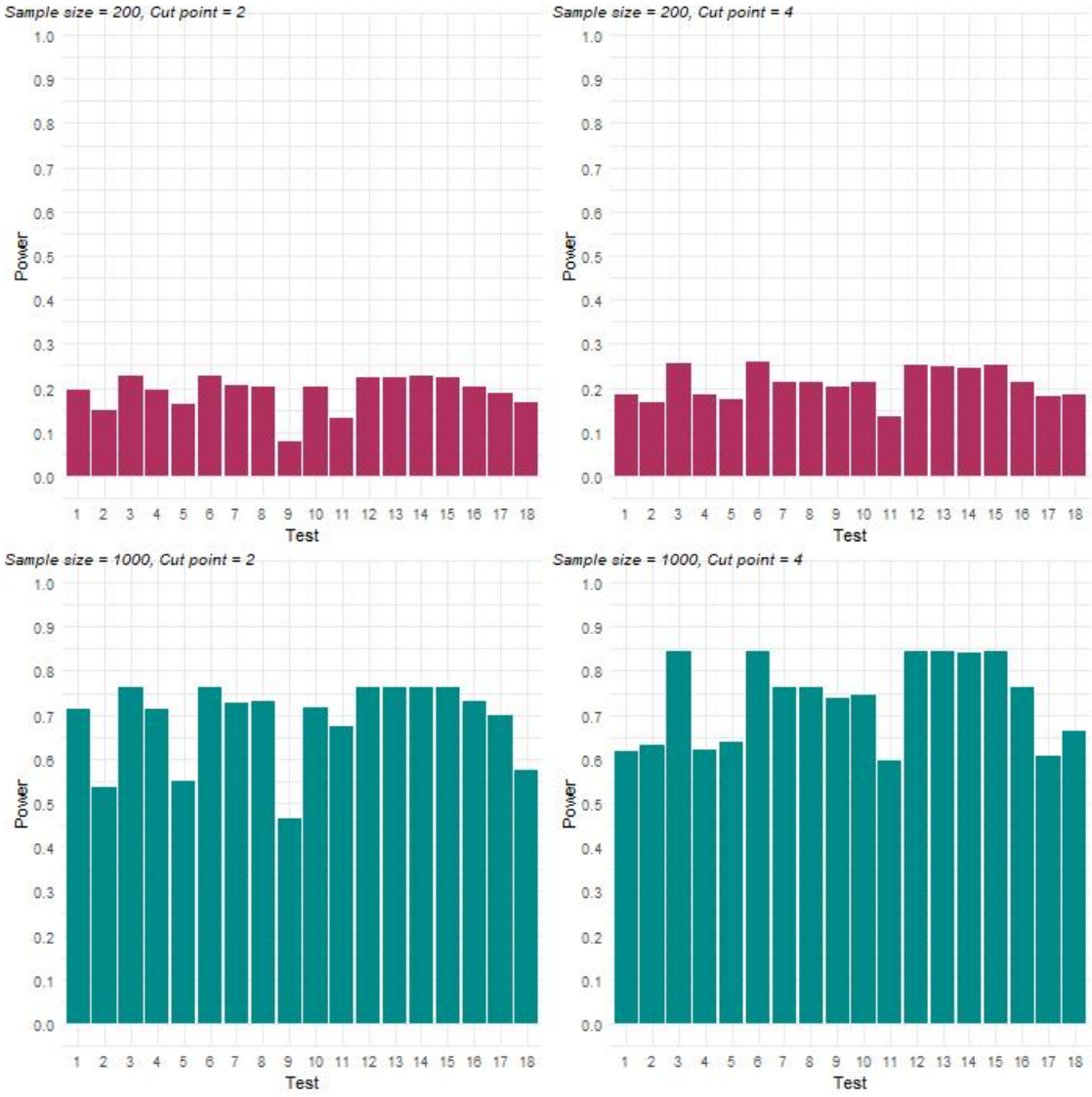

Figure A.6: Power of 18 tests for proportional hazards, for different sample sizes and cut points for the scenario of crossing hazards with initial HR = 1.2 and subsequent HR = 0.8.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	9.3	23.1	12.5	26.9
2	6.8	18.0	9.4	25.0
3	8.2	24.8	11.8	36.2
4	9.3	23.3	12.4	26.9
5	7.0	18.2	9.9	25.6
6	9.0	24.4	11.9	36.0
7	9.5	23.1	12.0	31.5
8	9.4	23.3	12.1	31.6
9	1.7	3.7	9.5	27.6
10	10.2	23.7	13.0	31.6
11	2.7	11.3	4.0	21.5
12	8.1	24.8	11.7	36.2
13	8.1	24.5	11.3	35.7
14	8.4	25.9	12.4	37.0
15	8.1	24.8	11.7	36.2
16	9.5	23.2	11.9	31.5
17	9.3	23.5	12.1	27.7
18	7.2	18.7	9.6	25.8

Table A.7: Power(%) of 18 tests for proportional hazards in the two-sample problem, for the scenario of long-term survivors with initial HR = 0.8 and subsequent HR = 0.8^2, for different sample sizes n and cut points 2 and 4.

Figure A.7: Power of 18 tests for proportional hazards, for different sample sizes and cut points for the scenario of long-term survivors with initial HR = 0.8 and subsequent HR = 0.8^2.

Appendix B

Simulation study B

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	24.8	87.8	48.3	99.5	68.8	100.0
2	24.7	88.1	47.1	99.5	68.1	100.0
3	25.6	88.8	48.3	99.5	66.6	100.0
4	25.6	90.7	45.8	99.4	65.7	100.0
5	25.9	89.2	47.8	99.6	65.0	100.0
6	32.3	92.6	56.6	99.9	68.5	100.0
7	6.9	16.6	18.0	56.1	40.3	94.7
8	20.8	72.2	42.2	97.6	67.8	100.0
9	20.7	72.2	41.8	97.6	67.4	100.0
10	27.8	91.1	45.6	99.6	59.5	100.0
11	27.8	91.1	45.6	99.6	59.5	100.0
12	20.4	71.8	41.8	97.6	67.4	100.0
13	20.8	71.7	43.9	97.9	68.8	100.0
14	20.2	71.8	44.3	98.3	69.1	100.0
15	29.5	91.2	49.1	99.6	67.6	100.0
16	28.5	90.5	46.0	99.6	64.0	100.0
17	21.6	74.5	45.6	98.3	69.8	100.0
18	26.3	89.8	46.4	99.2	66.3	100.0
19	31.2	92.7	56.7	99.9	69.0	100.0
20	19.6	69.0	39.8	97.2	65.7	99.9

Table B.1: Power(%) of 20 tests for treatment effect, for the early effect case with initial HR = 0.65 and subsequent HR ≈ 1 after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

Figure B.1: Power of 20 tests for treatment effect, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when an early effect with initial HR = 0.65 is observed.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	7.1	10.3	7.6	16.9	9.2	24.9
2	6.6	10.4	7.7	16.9	9.0	25.1
3	7.2	10.5	8.1	17.8	9.7	25.1
4	6.4	10.3	6.9	16.4	8.9	23.6
5	7.2	10.4	7.9	17.5	9.4	24.0
6	7.9	14.0	8.9	21.2	10.3	27.5
7	5.6	6.3	5.9	8.6	6.8	14.7
8	7.1	10.8	7.4	18.2	8.5	24.0
9	6.8	10.8	7.2	18.1	8.5	24.0
10	7.2	10.7	8.0	15.7	8.7	22.0
11	7.0	10.6	8.1	15.7	8.8	21.9
12	6.8	10.6	6.9	18.0	8.4	24.3
13	6.5	10.3	7.0	17.4	8.1	22.9
14	5.7	10.5	6.6	16.6	7.1	23.1
15	6.8	11.8	8.5	18.1	9.0	24.8
16	7.5	11.1	9.0	17.5	9.3	23.7
17	6.8	10.8	7.3	17.7	8.6	24.0
18	6.7	11.0	8.3	17.1	9.2	24.3
19	7.4	14.3	9.0	21.4	10.4	27.2
20	6.7	10.4	7.4	17.8	8.2	23.0

Table B.2: Power(%) of 20 tests for treatment effect, for the early effect case with initial HR = 0.9 and subsequent HR ≈ 1 after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

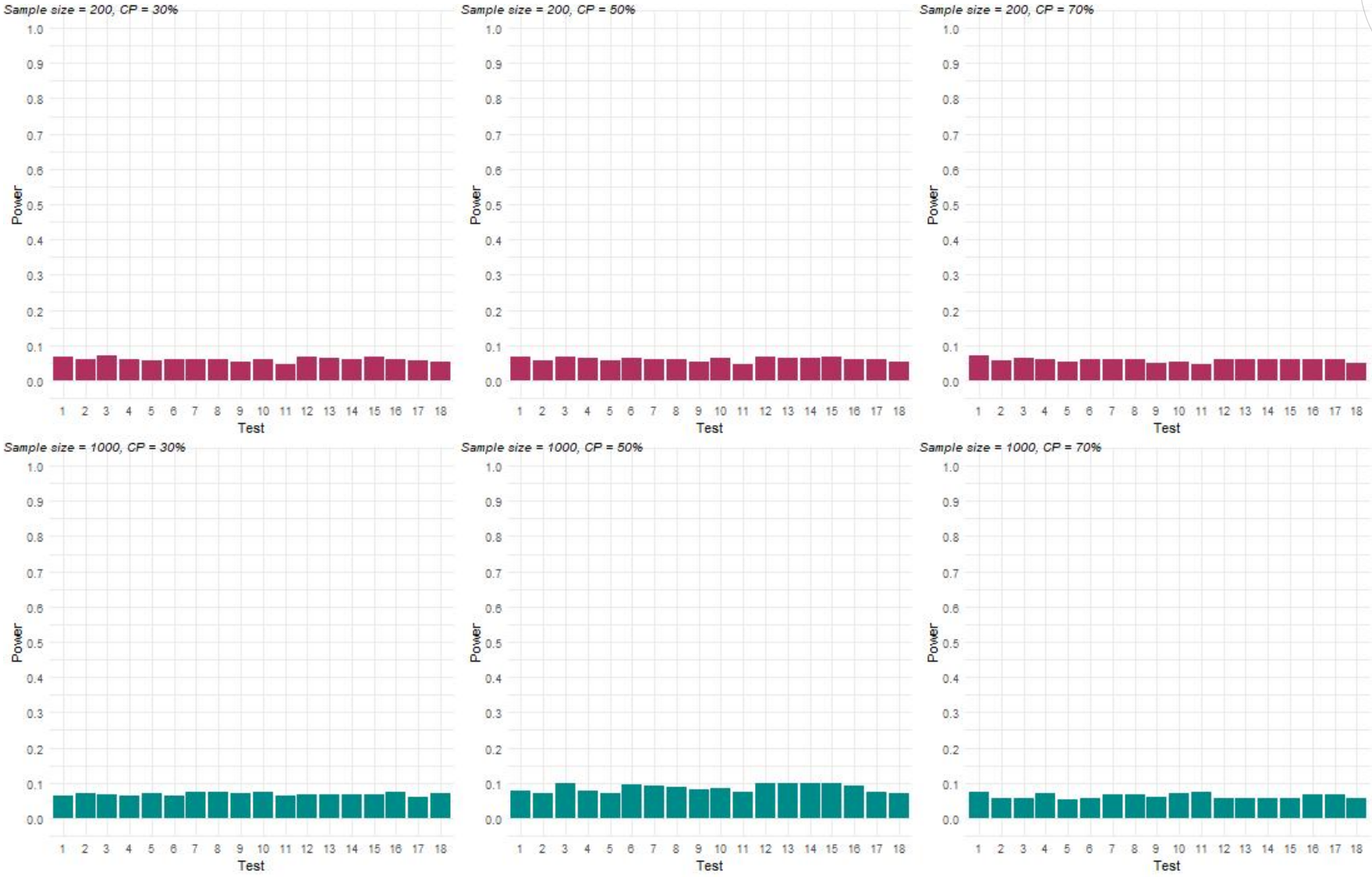

Figure B.2: Power of 20 tests for treatment effect, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when an early effect with initial HR = 0.9 is observed.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	59.0	100.0	36.2	96.8	14.6	67.4
2	59.2	100.0	36.8	96.7	15.5	68.4
3	59.3	99.8	37.1	97.1	16.5	70.5
4	56.6	100.0	39.2	96.9	19.9	78.4
5	58.8	99.8	37.7	97.1	16.5	71.7
6	23.0	78.8	11.0	32.0	7.6	9.8
7	64.5	99.9	47.5	98.6	22.2	79.5
8	50.9	99.3	26.9	88.6	11.8	45.2
9	50.7	99.3	26.9	88.5	11.6	45.2
10	54.9	99.9	39.4	97.4	18.7	75.1
11	54.9	99.9	39.2	97.4	18.7	75.1
12	49.8	99.3	25.9	88.2	11.0	44.5
13	54.0	99.3	29.8	92.7	14.1	56.8
14	46.8	99.3	23.9	90.9	9.3	50.3
15	43.9	99.1	22.8	84.4	10.2	40.4
16	61.3	99.9	41.3	97.9	19.4	74.7
17	52.5	99.3	28.9	91.7	13.2	52.5
18	54.1	100.0	37.3	97.8	19.4	78.1
19	24.8	80.7	11.4	34.1	7.5	9.9
20	52.1	99.5	28.3	90.4	12.7	50.2

Table B.3: Power(%) of 20 tests for treatment effect, for the late effect case with initial HR ≈ 1 and subsequent HR $= 0.65$ after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

Figure B.3: Power of 20 tests for treatment effect, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when a late effect with final HR = 0.65 is observed.

	Change point at x% of events					
	$x = 30$		$x = 50$		$x = 70$	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	8.6	23.2	7.2	16.2	5.5	9.7
2	8.5	22.7	7.0	16.2	5.6	9.9
3	8.5	23.0	7.3	16.9	6.3	10.3
4	7.9	22.1	7.2	15.4	6.2	9.6
5	8.5	21.6	7.4	16.8	6.5	10.4
6	7.5	12.6	7.6	7.5	6.8	5.8
7	8.3	26.0	7.4	20.5	6.6	12.7
8	7.9	21.3	7.1	15.7	6.4	9.6
9	7.7	21.2	7.0	15.7	6.3	9.6
10	8.1	21.5	7.1	16.1	6.5	9.9
11	8.1	21.5	7.1	16.0	6.5	9.9
12	7.4	21.1	6.9	15.6	6.4	9.7
13	8.0	21.9	6.7	16.5	6.7	10.0
14	6.3	21.1	6.1	15.1	5.7	9.6
15	8.7	18.9	7.1	11.9	6.2	7.6
16	10.1	24.2	8.3	16.9	7.0	11.0
17	8.0	21.8	6.7	15.7	6.5	10.1
18	8.7	21.9	7.2	16.0	6.7	10.5
19	7.3	12.5	7.6	7.6	6.7	5.8
20	8.0	21.9	7.2	16.9	6.0	10.3

Table B.4: Power(%) of 20 tests for treatment effect, for the late effect case with initial HR ≈ 1 and subsequent HR $= 0.9$ after 30%, 50% and 70% of events have been observed in the treatment group, for different sample sizes n.

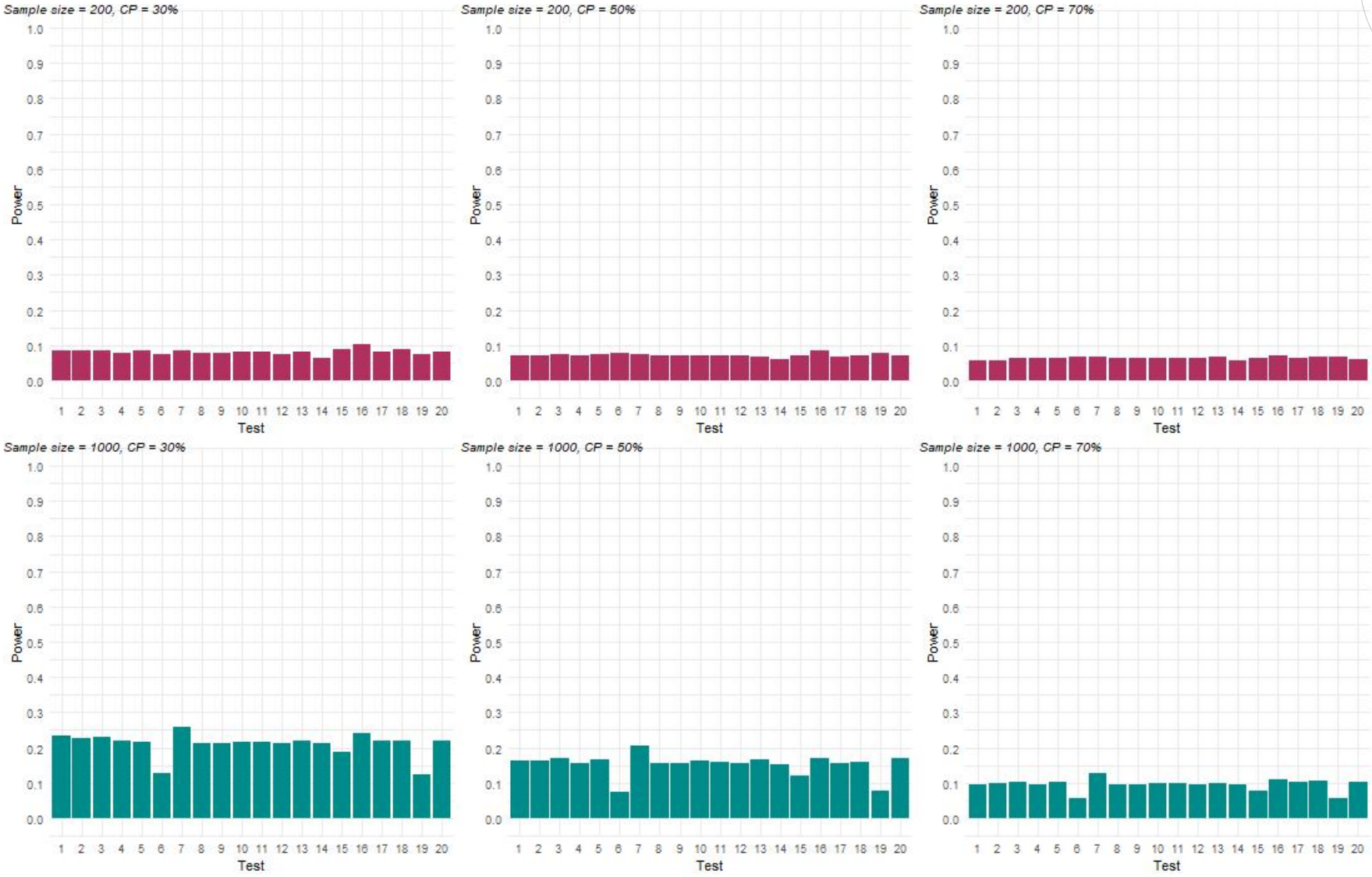

Figure B.4: Power of 20 tests for treatment effect, for two sample sizes and three change points (CP) at 30%, 50% and 70% of events in the treatment group, when a late effect with final HR = 0.9 is observed.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	26.0	90.1	24.9	89.3
2	25.5	90.0	25.0	89.6
3	26.7	90.6	26.2	90.6
4	28.6	93.7	28.1	93.6
5	27.5	90.8	26.9	90.9
6	33.7	94.0	33.4	93.7
7	4.6	8.1	6.0	6.2
8	18.7	71.3	16.4	59.0
9	18.4	71.3	16.2	59.0
10	33.3	96.5	34.0	95.8
11	33.2	96.5	34.0	95.9
12	18.3	71.2	16.3	59.1
13	29.4	88.9	18.9	65.7
14	30.8	89.3	19.2	67.2
15	32.5	95.6	32.0	95.4
16	32.5	95.6	32.0	95.8
17	29.6	89.4	19.7	69.6
18	33.8	96.8	30.6	95.3
19	32.6	93.5	32.1	93.2
20	16.7	68.5	15.0	54.0

Table B.5: Power(%) of 20 tests for treatment effect, for the scenario of crossing hazards with initial HR = 0.65 and subsequent HR = 1.10, for different sample sizes n and cut points 2 and 4.

Figure B.5: Power of 20 tests for treatment effect, for different sample sizes and cut points for the scenario of crossing hazards with initial HR = 0.65 and subsequent HR = 1.10.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	13.0	63.0	24.9	88.6
2	13.3	63.9	25.3	89.0
3	14.8	65.6	26.6	89.8
4	19.1	77.3	29.7	93.1
5	15.3	66.3	27.1	90.5
6	6.2	6.1	6.5	6.4
7	18.7	73.7	33.9	92.8
8	7.9	25.1	13.1	51.2
9	7.9	25.1	12.8	51.1
10	21.5	82.9	34.5	96.0
11	21.7	82.7	34.4	96.0
12	7.9	25.1	12.6	50.9
13	6.9	11.1	13.2	49.9
14	7.0	10.5	10.5	47.8
15	9.6	29.9	13.7	51.7
16	17.9	71.5	32.2	92.9
17	6.9	10.3	12.5	45.9
18	19.4	78.1	29.2	94.1
19	6.3	5.9	7.0	6.9
20	8.3	27.5	14.3	56.1

Table B.6: Power(%) of 20 tests for treatment effect, for the scenario of crossing hazards with initial HR = 1.10 and subsequent HR = 0.65, for different sample sizes n and cut points 2 and 4.

Figure B.6: Power of 20 tests for treatment effect, for different sample sizes and cut points for the scenario of crossing hazards with initial HR = 1.10 and subsequent HR = 0.65.

	Cut point			
	2		4	
Test	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
1	96.1	100	99.3	100
2	96.1	100	99.3	100
3	96.1	100	99.2	100
4	95.8	100	99.2	100
5	96.1	100	99.2	100
6	87.0	100	91.0	100
7	97.0	100	99.4	100
8	95.4	100	98.8	100
9	95.4	100	98.8	100
10	94.4	100	98.8	100
11	94.4	100	98.8	100
12	95.4	100	98.7	100
13	90.2	100	98.7	100
14	88.6	100	98.1	100
15	94.2	100	98.1	100
16	96.2	100	99.4	100
17	90.0	100	98.5	100
18	95.6	100	99.1	100
19	88.1	100	91.8	100
20	95.4	100	99.1	100

Table B.7: Power(%) of 20 tests for treatment effect, for the scenario of long-term survivors with initial HR = 0.65 and subsequent HR = 0.65^2, for different sample sizes n and cut points 2 and 4.

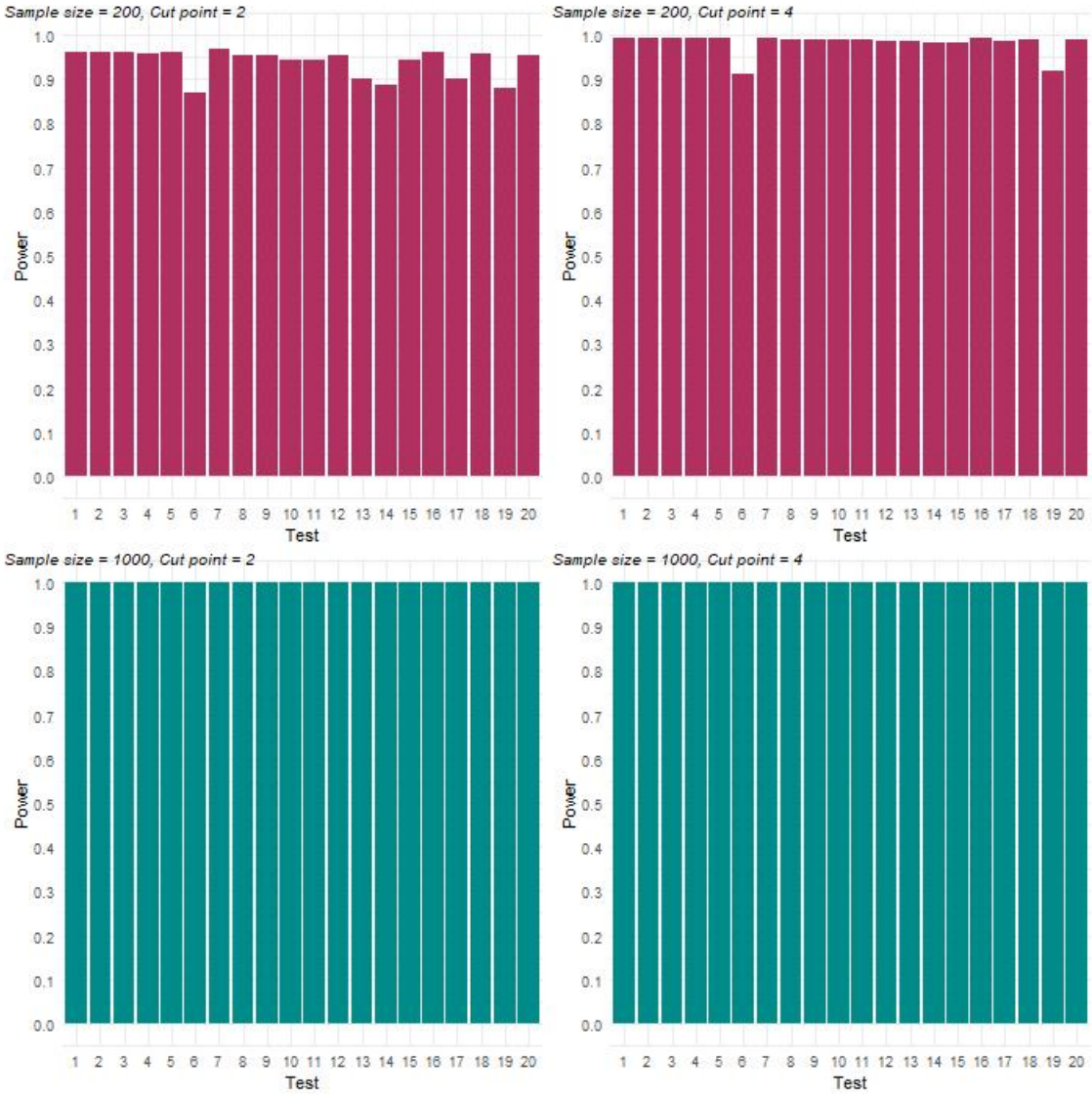

Figure B.7: Power of 20 tests for treatment effect, for different sample sizes and cut points for the scenario of long-term survivors with initial HR = 0.65 and subsequent HR = 0.65^2.

Appendix C

Simulated scenarios

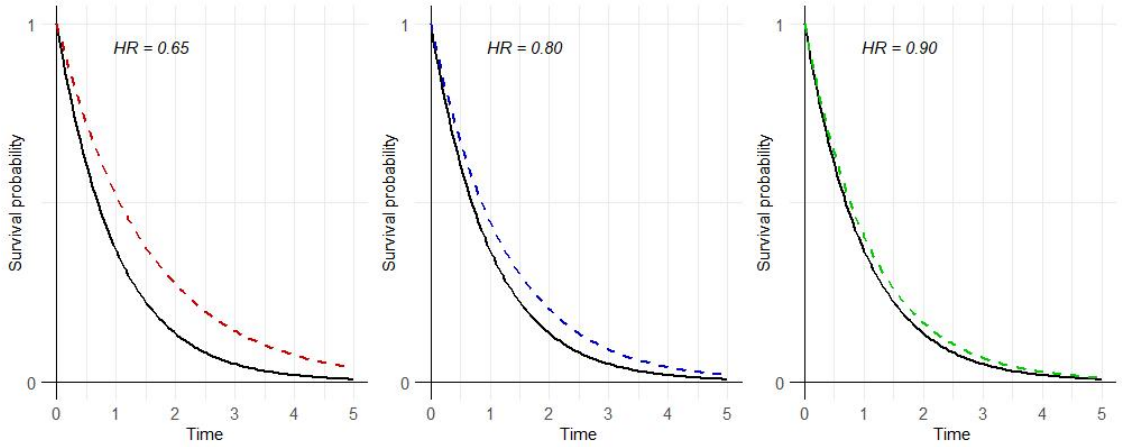

Figure C.1: Simulated scenarios for the case of proportional hazards, with baseline hazard equal to 1.

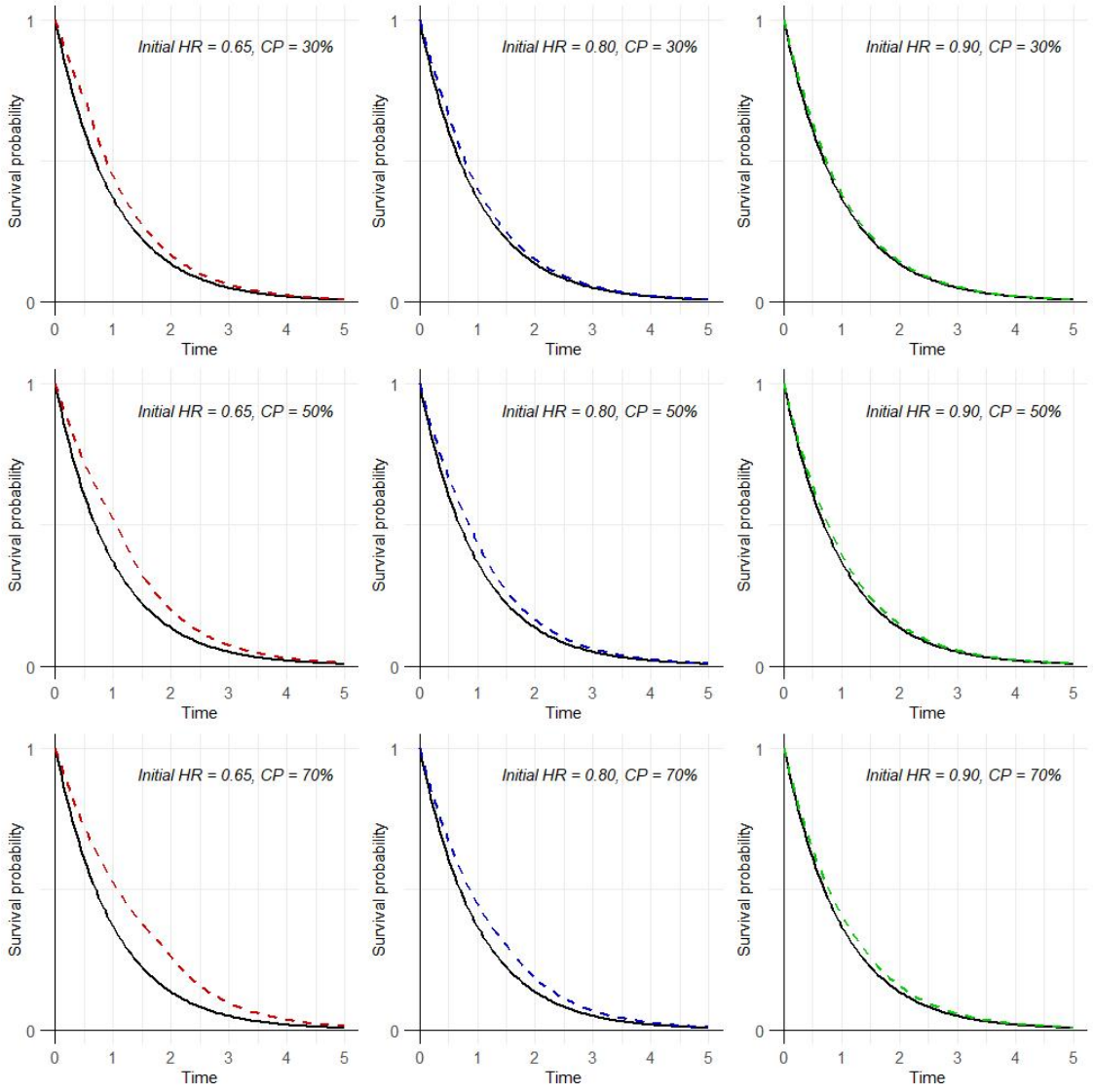

Figure C.2: Simulated scenarios for the early effect case with baseline hazard equal to 1, for three change points (CP) at 30%, 50% and 70% of events in the treatment group.

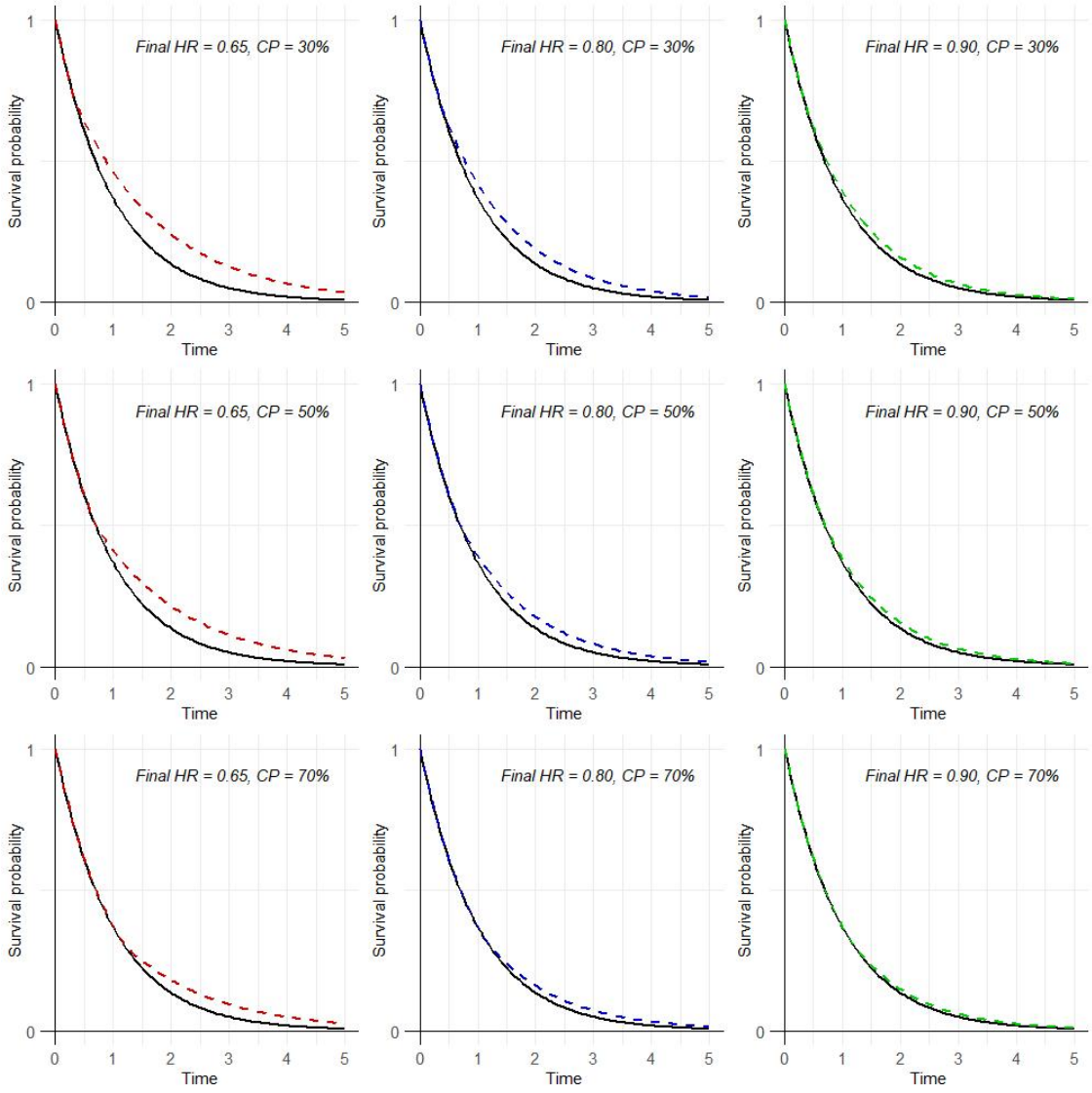

Figure C.3: Simulated scenarios for the late effect case with baseline hazard equal to 1, for three change points (CP) at 30%, 50% and 70% of events in the treatment group.

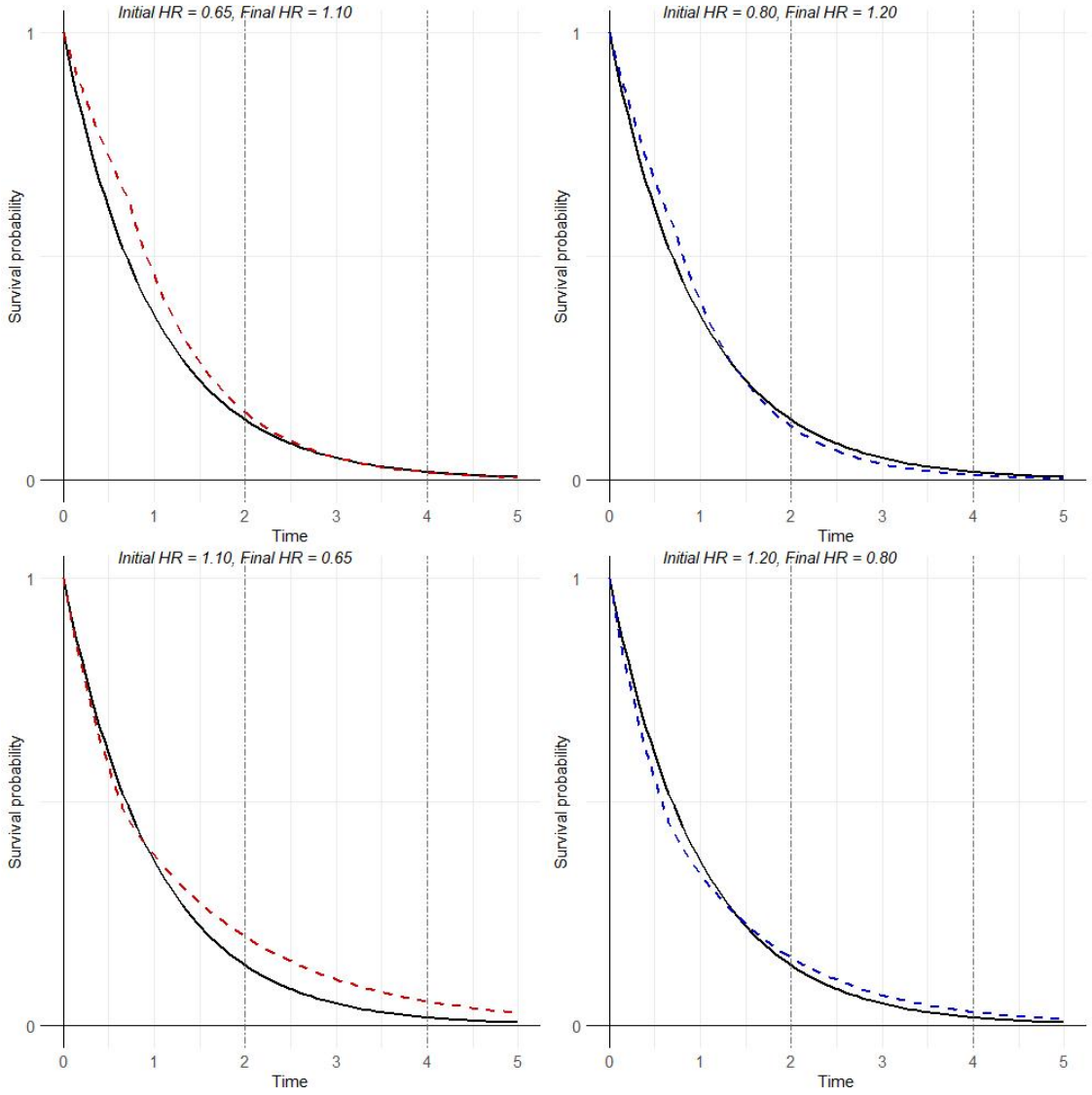

Figure C.4: Simulated scenarios for the crossing hazards case with baseline hazard equal to 1. The vertical dashed lines correspond to two pre-specified time cut points.

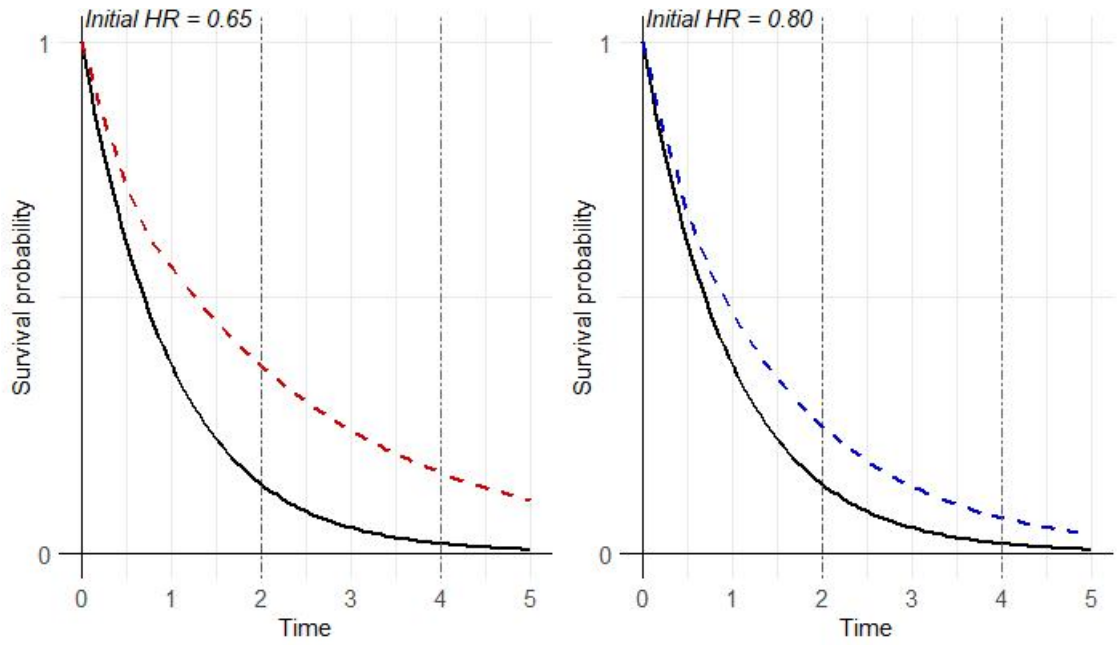

Figure C.5: Simulated scenarios for the case of long-term survivors with baseline hazard equal to 1. The vertical dashed lines correspond to two pre-specified time cut points.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Bibliography

Ananthakrishnan, R., S. Green, A. Previtali, R. Liu, D. Li, and M. LaValley (2021). Critical review of oncology clinical trial design under non-proportional hazards. *Critical Reviews in Oncology/Hematology 162*, 103350.

Andersen, P. K. (1982). Testing goodness of fit of Cox's regression and life model. *Biometrics*, 67–77.

Anderson, J. and A. Senthilselvan (1982). A two-step regression model for hazard functions. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 31*(1), 44–51.

Bagdonavičius, V., M. A. Hafdi, and M. Nikulin (2004). Analysis of survival data with cross-effects of survival functions. *Biostatistics 5*(3), 415–425.

Bagdonavičius, V. and R. Levulienė (2019). Testing proportional hazards for specified covariates. *Modern Stochastics: Theory and Applications 6*(2), 209–225.

Boyd, A. P., J. M. Kittelson, and D. L. Gillen (2012). Estimation of treatment effect under non-proportional hazards and conditionally independent censoring. *Statistics in Medicine 31*(28), 3504–3515.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89–99.

Breslow, N. E. (1972). Contribution to discussion of paper by DR Cox. *Journal of the Royal Statistical Society: Series B (Methodological) 34*, 216–217.

Breslow, N. E., L. Edler, and J. Berger (1984). A two-sample censored-data rank test for acceleration. *Biometrics*, 1049–1062.

Chappell, R. (1992). A note on linear rank tests and Gill and Schumacher's tests of proportionality. *Biometrika*, 199–201.

Chen, T.-T. (2013). Statistical issues and challenges in immuno-oncology. *Journal for immunotherapy of cancer 1*(1), 1–9.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(2), 187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika 62*(2), 269–276.

Dabrowska, D. M., K. A. Doksum, N. J. Feduska, R. Husing, and P. Neville (1992). Methods for comparing cumulative hazard functions in a semi-proportional hazard model. *Statistics in Medicine 11*(11), 1465–1476.

Dabrowska, D. M., K. A. Doksum, and J.-K. Song (1989). Graphical comparison of cumulative hazards for two populations. *Biometrika 76*(4), 763–773.

Dunkler, D., M. Ploner, M. Schemper, and G. Heinze (2018). Weighted Cox regression using the R package coxphw. *Journal of Statistical Software 84*, 1–26.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association 72*(359), 557–565.

Fleming, T. R. and D. P. Harrington (1991). *Counting processes and survival analysis*. John Wiley & Sons.

Fleming, T. R., D. P. Harrington, and M. O'sullivan (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. *Journal of the American Statistical Association 82*(397), 312–320.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika 52*(1-2), 203–224.

Gill, R. and M. Schumacher (1987). A simple test of the proportional hazards assumption. *Biometrika 74*(2), 289–300.

Good, I. and R. A. Gaskins (1971). Nonparametric roughness penalties for probability densities. *Biometrika 58*(2), 255–277.

Grambsch, P. M. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*(3), 515–526.

Hafdi, M. A. (2021). Testing the proportionality assumption for specified covariate in the Cox model. *Pakistan Journal of Statistics and Operation Research*, 435–449.

Harrell, F. E. and K. L. Lee (1986). Verifying assumptions of the Cox proportional hazards model. In *Proceedings of the eleventh annual SAS Users group international conference*, pp. 823–828. SAS Institute Inc, Cary, NC.

Harrington, D. (2014). Linear rank tests in survival analysis. *Wiley StatsRef: Statistics Reference Online*.

Harrington, D. P. and T. R. Fleming (1982). A class of rank test procedures for censored survival data. *Biometrika 69*(3), 553–566.

Hertz-Picciotto, I. and B. Rockhill (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, 1151–1156.

Irwin, J. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Epidemiology & Infection 47*(2), 188–189.

Kalbfleisch, J. D. and R. L. Prentice (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika 60*(2), 267–278.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association 53*(282), 457–481.

Karadeniz, P. G. and I. Ercan (2017). Examining tests for comparing survival curves with right censored data. *Stat Transit 18*(2), 311–28.

Karrison, T. G. (2016). Versatile tests for comparing survival curves based on weighted log-rank statistics. *The Stata Journal 16*(3), 678–690.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 26*(3), 227–237.

Kleinbaum, D. G., M. Klein, et al. (2012). *Survival analysis: a self-learning text*, Volume 3. Springer.

Kraus, D. (2007). Data-driven smooth tests of the proportional hazards assumption. *Lifetime Data Analysis 13*(1), 1–16.

Kvaløy, J. T. and L. Reiersølmoen Neef (2004). Tests for the proportional intensity assumption based on the score process. *Lifetime Data Analysis 10*(2), 139–157.

Lee, J. W. (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*, 721–725.

Lee, L. and R. Pirie (1981). A graphical method for comparing trends in series of events. *Communications in Statistics-Theory and Methods 10*(9), 827–848.

Lee, S.-H. (2007). On the versatility of the combination of the weighted log-rank statistics. *Computational Statistics & Data Analysis 51*(12), 6557–6564.

Lin, D. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *Journal of the American Statistical Association 86*(415), 725–728.

Lin, D. Y. and L.-J. Wei (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association 84*(408), 1074–1078.

Lin, D. Y., L.-J. Wei, and Z. Ying (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika 80*(3), 557–572.

Lin, R. S., J. Lin, S. Roychoudhury, K. M. Anderson, T. Hu, B. Huang, L. F. Leon, J. J. Liao, R. Liu, X. Luo, et al. (2020). Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. *Statistics in Biopharmaceutical Research 12*(2), 187–198.

Moreau, T., J. O'quigley, and M. Mesbah (1985). A global goodness-of-fit statistic for the proportional hazards model. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 34*(3), 212–218.

Nagelkerke, N., J. Oosting, and A. Hart (1984). A simple test for goodness of fit of Cox's proportional hazards model. *Biometrics*, 483–486.

Nguyen, V. Q. and D. L. Gillen (2012). Robust inference in discrete hazard models for randomized clinical trials. *Lifetime data analysis 18*(4), 446–469.

O'Quigley, J. and F. Pessione (1989). Score tests for homogeneity of regression effect in the proportional hazards model. *Biometrics*, 135–144.

Pena, E. A. (1998a). Smooth goodness-of-fit tests for composite hypothesis in hazard based models. *The Annals of Statistics 26*(5), 1935–1971.

Pena, E. A. (1998b). Smooth goodness-of-fit tests for the baseline hazard in Cox's proportional hazards model. *Journal of the American Statistical Association 93*(442), 673–692.

Pepe, M. S. and T. R. Fleming (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*, 497–507.

Pepe, M. S. and T. R. Fleming (1991). Weighted Kaplan-Meier statistics: Large sample and optimality considerations. *Journal of the Royal Statistical Society: Series B (Methodological) 53*(2), 341–352.

Peto, R. and J. Peto (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General) 135*(2), 185–198.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika 65*(1), 167–179.

Quantin, C., T. Moreau, B. Asselain, J. Maccario, and J. Lellouch (1996). A regression survival model for testing the proportional hazards hypothesis. *Biometrics*, 874–885.

Rao, C. R., C. R. Rao, M. Statistiker, C. R. Rao, and C. R. Rao (1973). *Linear statistical inference and its applications*, Volume 2. Wiley New York.

Royston, P. and M. K. Parmar (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine 21*(15), 2175–2197.

Royston, P. and M. K. Parmar (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine 30*(19), 2409–2421.

Royston, P. and M. K. Parmar (2014). An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials 15*(1), 1–10.

Royston, P. and M. K. Parmar (2016). Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Medical Research Methodology 16*(1), 1–13.

Rufibach, K. (2019). Treatment effect quantification for time-to-event endpoints–Estimands, analysis strategies, and beyond. *Pharmaceutical Statistics 18*(2), 145–165.

Sahoo, S. and D. Sengupta (2016). On graphical tests for proportionality of hazards in two samples. *Statistics in Medicine 35*(6), 942–956.

Sasieni, P. (1993). Maximum weighted partial likelihood estimators for the Cox model. *Journal of the American Statistical Association 88*(421), 144–152.

Schemper, M. (1992). Cox analysis of survival data with non-proportional hazard functions. *Journal of the Royal Statistical Society: Series D (The Statistician) 41*(4), 455–465.

Schemper, M., S. Wakounig, and G. Heinze (2009). The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine 28*(19), 2473–2489.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika 67*(1), 145–153.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika 69*(1), 239–241.

Sengupta, D., A. Bhattacharjee, and B. Rajeev (1998). Testing for the proportionality of hazards in two samples against the increasing cumulative hazard ratio alternative. *Scandinavian Journal of Statistics 25*(4), 637–647.

Senthilselvan, A. (1980). *Smooth estimates of the hazard function in Cox's regression approach and applications in cancer prognosis.* Ph. D. thesis, University of Newcastle upon Tyne.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*(318), 626–633.

Song, H. H. and S. Lee (2000). Comparison of goodness of fit tests for the Cox proportional hazards model. *Communications in Statistics-Simulation and Computation 29*(1), 187–206.

Struthers, C. A. and J. D. Kalbfleisch (1986). Misspecified proportional hazard models. *Biometrika 73*(2), 363–369.

Tarone, R. E. and J. Ware (1977). On distribution-free tests for equality of survival distributions. *Biometrika 64*(1), 156–160.

Therneau, T. M. and P. M. Grambsch (2000). The Cox model. In *Modeling Survival Data: Extending the Cox model*, pp. 39–77. Springer.

Therneau, T. M., P. M. Grambsch, and T. R. Fleming (1990). Martingale-based residuals for survival models. *Biometrika 77*(1), 147–160.

Wei, L. (1984). Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association 79*(387), 649–652.

Xu, R. and J. O'Quigley (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics 1*(4), 423–439.

Zhang, H., Q. Li, D. V. Mehrotra, and J. Shen (2021). CauchyCP: A powerful test under non-proportional hazards using Cauchy combination of change-point Cox regressions. *Statistical Methods in Medical Research 30*(11), 2447–2458.