Censoring Bias in Oncology Clinical Trials

Andrea Mattsson

A Bachelor's Degree project in Mathematical statistics in cooperation with the Danish Medicines Agency.

Supervisor from the Danish Medicines Agency: Aldana Rosso
Academic supervisor: Anna Lindgren
Examiner: Johan Lindström
Centre for Mathematical Sciences
Faculty of Science
Lund University, Sweden
Spring semester 2020

Abstract

Time to disease progression or death (PFS) is of main interest in oncology clinical trials. The data is commonly analyzed using a Cox model. Early discontinuation gives uncertainty in the estimated hazard ratio (HR). Censoring is generally used to handle discontinuations. However, if discontinuations are related to the patients' prognosis, the HR may be biased. By simulating PFS data, this bias was investigated. The direction and size of the bias is dependent on the proportion of censored patients in each treatment arm. Our results show that supplementary analyses with different strategies to handle discontinuation should be used to compare cancer treatments.

Keywords: survival analysis, censoring, progression-free survival, Cox proportional hazards model, oncology clinical trial, HER2CLIMB

Populärvetenskaplig sammanfattning

Cancer är en av våra vanligaste sjukdomar och det pågår mycket forskning kring att hitta nya behandlingar. För att en ny behandling ska kunna erbjudas till cancerpatienter behöver behandlingens effekt och säkerhet utvärderas och säkerställas i en patientstudie. Effekten bedöms vanligtvis genom att mäta tiden det tar för cancersjukdomen att sprida sig och/eller tills patienten avlider. Generellt får de deltagande patienterna antingen den nya behandlingen eller den etablerade behandlingen och sedan jämförs effekten av de olika behandlingarna.

Det händer att patienter avbryter sitt deltagande i studien, till exempel om de får svåra biverkningar eller om de byter till en annan behandling. Då uppstår en osäkerhet i beräkningen av behandlingens effekt eftersom dessa patienter inte fullföljde behandlingen som planerat. Vanligtvis antar man att de avhoppade patienternas tidsförlopp inte skiljer sig från de kvarvarande patienternas tidsförlopp, om de hade fortsatt delta i studien. Om det stämmer påverkas inte beräkningen av behandlingens effekt. Om däremot anledningen till patienternas avhopp är relaterad till deras sjukdomsprognos, till exempel om de inte tål den nya behandlingen eftersom de är mycket sjuka, blir beräkningarna felaktiga.

I den här uppsatsen undersöks storleken på det fel i jämförelse av tid till spridning eller död som uppstår när man gör antaganden om patienter som avbryter studien i förtid. Utöver det undersöks om felet gynnar eller missgynnar den nya behandlingen. Undersökningen gjordes genom att simulera data som efterliknar en nyligen genomförd bröstcancerstudie [1]. Resultaten visar att andelen av patienter som avbryter behandlingen i de olika behandlingarna har betydelse för åt vilket håll felet går. Ett exempel på det är om patienter som får den nya behandlingen avbryter i en större utsträckning än de som får den vanliga behandlingen. Om man i det fallet antar att de patienter som avbröt den nya behandlingen fick spridd cancersjukdom eller avled precis när de avbröt blir beräkningen av behandlingseffekten mer konservativ.

Acknowledgements

First of all, I would like to thank my supervisor, Aldana Rosso, Senior Adviser in Biostatistics at the Danish Medicines Agency, for her great support and guidance through each stage of this project. I am extremely grateful for her trust in me and for inspiring my interest in this exciting field of research. I would also like to thank the Danish Medicines Agency, especially the Department of Cancer Diseases, for offering me this opportunity and for contributing with encouragement and helpful comments. Lastly, I would like to thank my academic supervisor, Anna Lindgren at the Division of Mathematical Statistics, for the support and valuable comments on my work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In clinical research, clinical trials are a vital part of investigating new disease treatments, for instance cancer medicines. Besides the clinical considerations, statistics is an important tool, not only to make sure the trial is conducted in a valid way, but also to make sure that the data is interpreted correctly.

The objective with oncology clinical trials is to identify treatments which prolong patients' life and/or help to control the disease, while improving patients' quality of life. Therefore, time to progression of the disease and/or death are the most common variables used to compare between established and experimental treatments in the later phases of clinical development. Trials investigating potential cancer treatments take a long time, sometimes several years. Moreover, both experimental and established treatments may have serious side effects. Therefore, a large commitment is required from the patients who participate in clinical studies. In many cases, the study treatment may not be well tolerated by the patient and it will be discontinued earlier than planned. In other cases, the patients may not be willing to continue receiving the study treatment. Earlier discontinuations interfere with the statistical comparison between the experimental and established treatments and there are several ways to handle it in the statistical analysis. Generally, it is assumed that patients who discontinue treatment are similar to those who remained in the study and were observed until they experienced the event of interest (death or progression of disease) or the trial was terminated. Under this assumption, patients who discontinued the trial without experiencing the event are censored. However, if this assumption is incorrect the results will be biased [2][3]. Biased estimates can lead the regulatory authorities to granting approval of inefficient treatments and be the cause of, not only false hope for the patients, but also economical

risks. Censoring bias has been investigated previously, for instance in [2], [4] [5] and [6] However, the size and direction of the bias given by early discontinuation is still not fully understood. A sensitivity analysis in the form of simulations of different scenarios would give useful information about the validity of the results.

The objective of this thesis is to investigate the possible bias due to censoring of patients who discontinue early in phase 3 clinical trials. An estimate of the possible bias is computed through simulations of survival data based on a recent oncology clinical trial, HER2CLIMB [1]. This trial concerns an experimental treatment for human epidermal growth factor receptor 2 (HER2) positive breast cancer[1]. It investigates if adding tucatinib, a selective blocker of HER2, to an already established treatment combination (the control treatment), will increase the overall treatment effect. In this trial, the early discontinuation rate was around 30% (27.5% in the experimental treatment arm and 33,8 in the control treatment arm), which restricted the interpretation of the results. The simulations will first mimic the results from the HER2CLIMB trial and then be modified in order to investigate the possible bias from different censoring scenarios.

[1]HER2 is a protein that is involved in cell growth. An overexpression of HER2 (i.e. HER2 positive breast cancer) leads to faster cancer cell growth and faster spread of cancer [7], i.e. a more aggressive cancer.

Chapter 2

Background theory

2.1 Clinical trials

A clinical trial is an experiment which investigates the safety and efficacy of an experimental treatment for some disease using subjects. Typically, the subjects are patients that are treated for the disease of interest. There are four phases during the development program of a new potential cancer treatment. Phase I and II mainly focus on the safety of the new drug, such as potential adverse reactions, for example diarrhea. The optimal dose level is also investigated. If the drug seems to show a potential treatment effect with a manageable safety profile, a larger scale phase III trial is conducted. In a phase III trial the effect of the new therapy is generally compared to a control therapy which is considered standard of care, i.e. what is currently used as treatment in most cases. Patients are assigned into treatment arms through a randomization process. This means that the patient is assigned to a treatment arm at random, without knowing before-hand which treatment she/he will receive. The main purpose for randomization is to be able to compare the arms in a statistically valid manner. The randomization process also makes the adjusting for other factors less important, since the groups of patients are comparable. In order to minimize bias, the study can be blinded, or even double blinded. This means that the patient does not know which treatment arm she/he belongs to (single-blind). When double blinding is applied, neither the patient nor the investigator or other personnel involved in the study know which treatment arm the patient belongs to. Phase IV investigates the efficacy and safety of the experimental treatment further, mostly after the market approval [8].

Before the start of the clinical trial, a protocol that specifies the study design and plan for the

analysis must be developed. In the protocol, the research questions, i.e. the objectives, to be addressed are presented, as well as a plan for how patients will be treated, how the treatments will be compared and which data that will be collected. It describes both the efficacy and safety objectives of the trial. The protocol needs to fulfil the principles of Good Clinical Practice and the study should comply with the Helsinki declaration [9].

A clinical trial has primary objectives, which are the most important ones, and secondary objectives, which are used to support the primary objectives. The primary variables, or endpoints, of the study should address the primary objective of the study. Secondary variables are related to the secondary objectives. In addition to the protocol, a statistical analysis plan (SAP) is developed to specify how the statistical analysis will be performed. For instance, which statistical method will be used to analyze the variables collected during the study, how to deal with missing data, and the randomization process are specified in the SAP [10]. Efficacy is the measured effect of the new treatment in the clinical trial. The treatments are compared using several measures or variables, for example length of survival or time to disease progression [8]. If efficacy is an objective of the trial, it also must examine possible side effects or harms. This is equally important as the efficacy objective [11].

In oncology clinical trials there are different endpoints of interest, such as overall survival (OS) and progression-free survival (PFS). OS measures the time from randomization until death, due to any cause, and is precisely measured and easy to track. A disadvantage is that it also includes deaths that are not directly connected to cancer. PFS measures the time until the cancer has progressed, or death, which of the two events that comes first. Progression of cancer means that the disease has advanced or spread. It is important that the definition of the progression is clearly stated ahead of the study, in order to decrease potential bias. A standardized scale, called RECIST, is typically used in solid tumors, to assess progression of the disease [12]. The advantages of PFS as an endpoint is that it needs a smaller sample size compared to OS since this is a more frequent event than OS [13]. Also, PFS generally needs a shorter follow-up period than OS, depending on how fast the disease advances. A disadvantage of PFS is that it is harder to track since progression can only be discovered when the patient is in a follow-up visit at the clinic, for example via a scan. Another disadvantage is that it does not necessarily correlate with the survival and this could be a problem if OS is an important objective of the study [14] [15].

2.2 Survival analysis

Survival analysis, also called time-to-event analysis, studies the time until an event. The reason for using survival analysis in cancer clinical trials is that the efficacy of an experimental treatment is studied by measuring time to progression or death (PFS time). In this thesis, standard survival analysis methods, such as Kaplan-Meier and Cox models, will be used to analyse PFS.

2.2.1 The survival function

In survival analysis, the main focus is the time T until some event. Therefore, it is convenient to use the survival function $S(t)$, which is defined as:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t),$$

where $F(t)$ is the cumulative distribution function of T.

2.2.2 The hazard function

An even more convenient way to model the survival is by the hazard function. It is described as the instantaneous failure rate and is the probability of failure in an infinitesimal time interval τ, given that you have survived up to now. It is defined as:

$$h(t) = \lim_{\tau \to 0} \frac{P(t < T < t + \tau | T > t)}{\tau}.$$

The hazard function is related to the survival function by

$$S(t) = \exp\left(- \int_0^t h(u) du \right).$$

This means, in theory, that if one of them is known, then so is the other [3].
The cumulative hazard function $H(t)$ is defined through the hazard function as:

$$H(t) = \int_0^t h(u) du.$$

2.2.3 Censoring

Censoring is when an event of interest, for example disease progression or death, is not observed, meaning that incomplete information is available about the time-to-event of some subjects. In clinical trials, right-censoring is the most common type. This is when the event happened after the last observed time, i.e. the true survival time is longer than the observed survival time. There is also left-censoring, when the true survival time is shorter than the observed time, and interval censoring, when the true survival time only is known to be within a time interval [3]. This thesis will only consider right-censoring. In figure 2.1, different types of right-censoring are depicted in a graph.

There are different types of censoring assumptions. Random censoring is when the censored subjects are assumed to have the same risk profile as the remaining subjects in the risk set. Hence, the censored patients are assumed to be represented by all the remaining patients in the trial. In this case, no bias is expected since no information is lost. On the other hand, non-random censoring indicates that censored subjects do not have a similar risk as the subjects remaining in the study. This means that it is not sufficient to only consider the observed events to estimate the survival of the censored subjects. Independent censoring, as opposed to dependent censoring, is random but only within a subgroup of the whole population, i.e. random censoring conditioned on some covariates. Hence, the censored patients are assumed to be represented by all the remaining patients within its subgroup. The assumption of independent censoring is weaker and less restrictive than the random censoring assumption. Non-informative censoring, as opposed to informative censoring, is when the distribution of the event times T gives no information about the distribution of the censoring times C [3].

The assumption of independent and non-informative censoring are often, but not necessarily, valid at the same time, because they are not equivalent to each other. In the same way, dependent censoring can imply informative censoring and vice versa, but not always. See figure 2.2 for a presentation of the relation between the different types of censoring assumptions, illustrated in a Venn diagram.

All methods and tests presented in this section rely on the assumption of independent censoring. This means that if the censoring is not independent, the validity of the model is compromised [3].

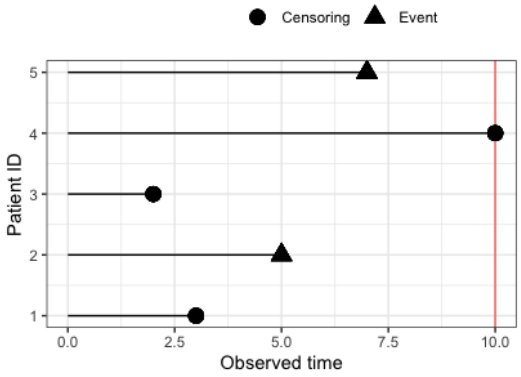


Figure 2.1: Graph of event and censoring times for five subjects. The circles denote that a patient was censored. The triangles denote that a patient experienced an event. Subject 2 and 5 experience the event at times 5 and 7 respectively. Subject 1 and 3 are right-censored at time 3 and 2 respectively, due to discontinuation. Subject 4 is administratively censored at time 10 since this is the end of the study.

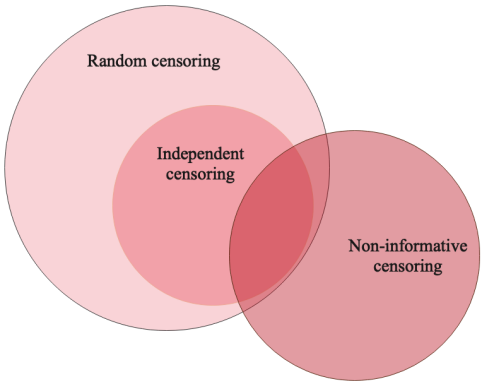


Figure 2.2: A Venn diagram that describes the relation between the different types of censoring assumptions. Independent censoring implies random censoring but not necessarily the other way around. All three may be valid at the same time but not by definition. The same relations are valid for the counterparts, i.e. non-random, dependent and informative censoring.

In clinical trials, censoring may occur if patients are lost to follow-up or if the patient has not yet experienced the event when the follow-up period is ended. The latter is called administrative censoring and is considered to be random since the end of the study date is not related to the patients' PFS time [16].

Due to censoring, the observed endpoint of a patient is controlled by two time processes. The PFS event time is denoted by the random variable $E \sim F_E(t)$ and the censoring time is denoted by $C \sim F_C(t)$, where $F_E(t)$ and $F_C(t)$ are the PFS event time and censoring time distributions respectively. The observed time, T, is then the time that happened first, i.e. the minimum of these, and is defined as

$$T = \min(E, C) \tag{2.1}$$

A censoring variable δ_i indicates if the patient is censored or not, i.e. denotes which of the PFS event or censoring happened first. It is defined as

$$\delta_i = \begin{cases} 0 & \text{if } T = C \\ 1 & \text{if } T = E \end{cases}$$

Assume that a clinical trial consist of n patients. Each patient i is then represented by a triplet $(T_i, \delta_i, \mathbf{x}_i)$, for $i = 1, \ldots, n$. The observed endpoint for patient i, T_i is i.i.d T, and δ_i indicates if the patient was censored or not. The vector $\mathbf{x_i} = (x_{1i}, \ldots x_{di}) =$ stores the values of the covariates for each patient i. In this thesis only one covariate, x_i, will be considered. It indicates which treatment arm patient i belongs to and is defined as

$$x_i = \begin{cases} 0 & \text{if patient } i \text{ is in the control treatment arm.} \\ 1 & \text{if patient } i \text{ is in the experimental treatment arm.} \end{cases}$$

Censoring does not necessarily lead to biased estimates. If it is independent or non-informative, the censored subjects are considered to be represented by the remaining subjects in the risk set, hence their hazard is taken into account. However, if the censoring is related to the survival function, for instance if the patients at higher risk tend to discontinue the trial before progression or death, this is not the case and bias may occur.

2.2.4 The Kaplan-Meier estimator

The Kaplan-Meier estimator, also called the product limit estimator, is a non-parametric estimator of the survival function. The estimation is computed stepwise for each interval between the ordered event times, and results in a right-continuous step function. It estimates the probability of surviving past an event time by computing the probability of surviving the previous event time multiplied by the conditional probability of surviving past the next event time given that you have survived up to the event time. For each step an interval is defined as $I_{(f)} = [t_{(f)}, t_{(f)})$ where $t_{(f)}$ is an ordered event time, for $f : t_{(f)} < t$, for a given time point t. The number of events at each ordered event time is denoted $d_{(f)}$. The number of subjects for which the event has not yet occurred is denoted $R_{(f)}$ and is called the risk set. The estimation takes censoring into account by not counting the censored subjects into the risk set after censoring but neither counting them as an event. In that way the Kaplan-Meier estimator uses the information about the censored subject just until they are censored.

The estimated survival probability in a given time point $t \in I_{(f)}$ is then

$$\hat{S}_{KM}(t) = \hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \cdot \frac{R_{(f)} - d_{(f)}}{R_{(f)}} = \prod_{k:\, t_{(f)} < t} \frac{R_{(k)} - d_{(k)}}{R_{(k)}}$$

Two-sided 95% confidence intervals for the Kaplan-Meier estimated survival curve are obtained by either computing the variance of the estimate by the frequently used Greenwood's formula [17] [3] [18] or by the delta method [17]. The estimator is asymptotically normally distributed and therefore the standard normal distribution p-quantile λ_p is used.

$$C.I._{KM} = \hat{S}_{KM}(t) \pm \lambda_{0.975} \sqrt{\mathrm{Var}(\hat{S}_{KM}(t))}$$

The assumptions made for the Kaplan-Meier estimator is that the censoring is independent and that the event and censoring times are exact. If the assumptions are not met, bias may occur [8].

2.2.4.1 Median survival time

The median survival time is defined as the time t such that $S(t) = \frac{1}{2}$. Since Kaplan-Meier survival curves are step functions it is not necessarily continuous at $t = \frac{1}{2}$, and a more general definition for the median survival is

$$t_{median} = \inf\{t : S(t) \leq \frac{1}{2}\}$$

The confidence interval for the median survival time is given by the inequality

$$(\hat{S}_{KM}(t) - \frac{1}{2})^2 < \chi_1^2 \cdot Var(\hat{S}_{KM}(t))$$

In clinical trials, this is the time when 50% of the study subjects have experienced the event. This is a descriptive statistical measure that can be used to support the comparison the survival of two different groups, for instance two treatment arms in a clinical trial. It can be estimated via the Kaplan-Meier estimate of the survival function [8]. In this thesis, the PFS median time will be considered, since PFS is the primary endpoint.

2.2.4.2 The Log-rank test

In order to compare two survival curves for two different groups and test if the observed difference is statistically significant, the log-rank test can be used. It is a non-parametric

large sample chi-square test. The null hypothesis H_0 is that the curves are overall equivalent to each other. The log-rank test compares the observed number of events and the expected number of events under the null hypothesis, for each event time [17].

In this thesis two groups, group 0 (experimental treatment arm) and group 1 (control treatment arm), will be compared to each other and therefore the test is defined for comparison of two groups.

In order to define the test statistic, some further definitions about status at each event time need to be specified. Let $n_{(1f)}$ be the number at risk for group 1 and let $d_{(1f)}$ be the number of events for each event time $t_{(f)}$. It can be shown that $d_{(1f)}|n_{(0f)},\ n_{(1f)},\ d_{(f)}$ follows the hypergeometric distribution [16]. This gives that the expected number of events $e_{(1f)}$ at event time $t_{(f)}$ is

$$e_{(1f)} = \mathrm{E}(d_{(1f)}) = \frac{n_{(1f)}d_{(f)}}{n_{(f)}},$$

and the variance $v_{(1f)}$ is

$$v_{(1f)} = \mathrm{Var}(d_{(1f)}) = \frac{n_{(0f)}n_{(1f)}d_{(f)}(n_{(f)} - d_{(f)})}{n_{(f)}^2(n_{(f)} - 1)}$$

The test statistic S is then computed as,

$$S = \frac{(\sum_{f=1}^{n}(d_{(1f)} - e_{(1f)}))^2}{\sum_{f=1}^{n} v_{(1f)}}$$

.
The computations are symmetric if group 2 is considered instead, resulting in the same statistic.
Under H_0, $S \sim \chi_\alpha^2(1)$ where α is the significance level.

2.2.5 The Cox proportional hazards regression model

The Cox model is a semi-parametric regression model that can be used to compare two groups with different treatments, for instance a new treatment compared to a control treatment. It estimates the relative difference in treatment effect.

The general form of the Cox model is expressed in the hazard function $h(t, \mathbf{x_i})$ for patient i

with covariate vector $\mathbf{x_i} = (x_{1i}, \ldots x_{di})$ and regression coefficients $\beta = (\beta_1, \ldots, \beta_d)$ at time t and is given by

$$h(t, \mathbf{x_i}) = h_0(t) \cdot \exp\left(\sum_{j=1}^{d} \beta_j x_{ij}\right)$$

It is a semi-parametric regression model since the baseline hazard function, $h_0(t)$, is undefined. This makes it more robust than a fully parametric model, for instance the Weibull hazard model. However, as a consequence, the model cannot be used to predict unless the baseline hazard is defined [3].

The baseline hazard is dependent on the time variable but the explanatory variables are assumed to be time-independent for the standard Cox model. If they are considered to be time-dependent, the extended Cox model can be used [3].

2.2.5.1 Adjusted survival curves

The survival curves can be estimated by the Cox model by using the estimated regression coefficients and specify some value for each covariate. The adjusted survival curves are then defined as

$$\widehat{S}(t, \mathbf{x}) = [S_0(t)]^{\exp \sum_{j=1}^{d} \beta_j x_{ij}}$$

where $S_0(t)$ is the baseline survival function, defined by the baseline hazard function.

The adjusted survival curves are not the same as the Kaplan-Meier curves since the latter do not adjust for covariates.

2.2.5.2 The hazard ratio and proportional hazards

The relative measure of difference in treatment efficacy between two groups is estimated by the hazard ratio (HR). The hazard ratio is a fraction of two hazards for two groups with different covariate values. Assume that two groups of patients are to be compared, group 0 with population size n_0 and group 1 with population size n_1. Group 0 has covariate values $\mathbf{x_i} = (x_{1i}, \ldots x_{di})$ for each patient i, $i = 1, \ldots, n_0$ and group 1 with covariate values $\mathbf{x_i^*} = (x_{1i}^*, \ldots x_{di}^*)$ for each patient i^*, $i^* = 1, \ldots, n_1$. Given the estimated regression coefficients $\hat{\beta} = (\beta_1, \ldots, \beta_d)$, the hazard ratio between patient i in group 0 and patient i^* in group 1 is

then computed by

$$HR = \frac{h_0(t)\exp\left(\sum_{j=1}^{d}\hat{\beta}_j x_{ij}\right)}{h_0(t)\exp\left(\sum_{j=1}^{d}\hat{\beta}_j x^*_{ij}\right)} = \exp\left(\sum_{j=1}^{d}\hat{\beta}_j(x_{ij}-x^*_{ij})\right)$$

If there is only one binary covariate with estimated regression coefficient $\hat{\beta}$, then the hazard ratio becomes $HR = \exp(\hat{\beta})$ between the two groups. This is the case in this thesis, where the only covariate is the treatment indicator. If $HR < 1$, the hazard for group 0 is estimated to be smaller than the hazard for group 1. If $HR = 1$, the hazards of the two groups are estimated to be equal.

The baseline hazard $h_0(t)$ does not need to be specified since it is cancelled out in the fraction when the hazard ratio is assumed to be constant over time. This is the main assumption for the Cox model and is called proportional hazards (PH). If the PH assumption is met can be evaluated graphically, by goodness of fit tests, or by fitting a model with time-dependent variables [3].One way of evaluating graphically is to use the log-negative-log-plot. Let's assume that we have two estimated survival curves, adjusted for treatment group, $\widehat{S}_1(t)$ and $\widehat{S}_2(t)$. If the PH-assumption is met, then the plotted curves of $\ln(-\ln(\widehat{S}_1(t))$ and $\ln(-\ln(\widehat{S}_2(t))$ should be parallel. If they cross, diverge or converge, the PH-assumption is violated [3] [17].

2.2.5.3 The Cox partial likelihood

The Cox likelihood is a partial likelihood which is defined based on the order of event and censoring times, instead of the survival distribution. This is due to the fact that the survival distribution is not completely defined, due to the unknown baseline hazard function $h_0(t)$. The assumptions that are made are the same as those described previously, i.e. that we know the event and censoring times exactly and that the censoring is independent.

The Cox model is fitted using maximum likelihood estimation (MLE). However, in order to handle censored data the estimation needs to be adapted , so that all information until censoring is used in the model. The full likelihood is defined by the probability density functions. The probability for the subjects who experiences the event is used but for the censored

subjects the survival function is used instead. This allows the model to adapt to the fact that we know that the survival times of the censored subjects are *at least* the censoring time but perhaps greater.

Consider the triplets $(T,0,\mathbf{x_i})$ and $(T,1,\mathbf{x_i})$ separately for a patient i. For $(T,0,\mathbf{x_i})$ the PFS time is at least $t_i = T$, since the censor variable δ_i is equal to 0. Hence the probability for PFS is given by $S(T,\beta,\mathbf{x_i})$. For $(T,1,\mathbf{x_i})$ on the other hand, the PFS time is exactly $t_i = T$, hence the probability for PFS is given by the p.d.f. $f(T,\beta,x_i)$. For a sample of m independent observations $(t_i,\delta_i,\mathbf{x_i})$, $i = 1,\ldots,m$, the likelihood function $L(\beta)$ is defined in [17] as

$$L(\beta) = \prod_{i=1}^{m} [f(t_i,\beta,\mathbf{x_i})]^{\delta_i} [S(t_i,\beta,\mathbf{x_i})]^{1-\delta_i}$$

For simplicity, the log-likelihood function $l(\beta)$ is maximized with respect to β and this is equivalent to maximizing $L(\beta)$ since the log-transform is monotone.

$$l(\beta) = \sum_{i=1}^{m} \left(\delta_i \log\left(f(t_i,\beta,\mathbf{x_i})\right) + (1-\delta_i)\log\left(S(t_i,\beta,\mathbf{x_i})\right)\right)$$

Using that $f(t_i,\beta,\mathbf{x_i}) = h(t_i,\beta,\mathbf{x_i}) \cdot S(t_i,\beta,\mathbf{x_i})$, this is equivalent to

$$l(\beta) = \sum_{i=1}^{m} \left(\delta_i \log\left(h_0(t_i) + \delta_i \mathbf{x_i}\beta + e^{\mathbf{x_i}\beta}\log\left(S_0(t_i)\right)\right)\right)$$

However, since the baseline hazard and survival functions are unknown, the partial log-likelihood is defined as

$$l_p(\beta) = \prod_{i=1}^{m} \left(\frac{e^{\mathbf{x_i}\beta}}{\sum_{i=1}^{m} e^{\mathbf{x_i}\beta}}\right)^{\delta_i}.$$

The partial likelihood was defined in [19] in 1972 and is shown to have the same distributional properties as the full likelihood, but without having to define the baseline hazard [20]. It can be used for log-likelihood ratio, score and the Wald tests, using the asymptotic properties of MLE. Furthermore, if the proportional hazards assumption is met, the score statistic is equivalent to the log-rank statistic, defined in section 2.2.4.2, hence only the log-rank test will be used in this thesis.

2.2.5.4 Cox regression compared to other regression models

Similarly to linear and logistic regression a model is estimated when using Cox regression. In linear regression the outcome variable of interest is some continuous variable and the measure of effect is the regression coefficient. In logistic regression the outcome variable of interest is a binary outcome and the measure of effect is the odds ratio expressed as the exponential of the regression coefficient. In survival analysis, the variable is the time until an event and the measure of effect is also expressed as the exponential of the regression coefficient but in this case is equal to the hazard ratio [3].

The main difference between Cox regression and linear or logistic regression is that it can handle censoring [16]. In linear and logistic regression the follow-up time is not tracked and therefore these models do not register if a patient did not complete an observation period [3].

2.2.6 Parametric survival distributions

There are several distributions that are suitable for simulations and inference of survival data. Since the endpoint variable is a measure of time, the distribution should preferably be non-negative. If a parametric distribution is assumed, the task is to choose or fit appropriate parameter values, depending on whether the goal is to simulate or estimate.

Previously, the hazard function was shown to be equivalent to the survival function. It is also equivalent to the density function $f(t)$ via the relation

$$f(t) = h(t)\exp(-H(t)).$$

If the hazard function is constant over time, the survival distribution is exponential. If this is clearly not the case, there are other suitable survival distributions such as Gamma, Log-normal and Log-logistic. Below, the Weibull distribution is chosen to simulate the event data and the exponential distribution is chosen to simulate the censoring data [21] [22].

2.2.6.1 The Weibull distribution

The Weibull distribution is a suitable choice for simulating or model survival data, since it is easy to define the parameters such that the median survival times for both treatment groups attain the wanted values. It is non-negative, continuous and its survival function is defined as

$$S(t) = \exp\left(-(\rho t)^{\lambda}\right)$$

where ρ is the scale parameter and λ is the shape parameter [23]. The hazard function then becomes

$$h(t) = \lambda\rho(\rho t)^{\lambda-1}$$

In order to obtain the wanted median survival time t_{median}, for a fixed ρ, λ is given by

$$\lambda = \frac{-\log(1/2)}{t_{median}^{\rho}} \tag{2.2}$$

Chapter 3

The HER2CLIMB study

3.1 Tucatinib in combination with trastuzumab and capecitabine in previously treated patients with HER2-positive metastatic breast cancer

Tucatinib is an investigational oral selective blocker for a receptor called human epidermal growth factor receptor 2 (HER2) tyrosine kinase that is present in one breast cancer subtype, HER2-positive breast cancer. Trastuzumab in combination with capecitabine is already an established treatment for this type of cancer. The HER2CLIMB trial investigated if the addition of tucatinib in combination with trastuzumab and capecitabine delays disease progression and/or improves survival in previously treated patients with HER2-positive metastatic breast cancer [1]. The primary and most important secondary endpoints of the study are PFS and OS respectively. Since censoring mostly affects PFS more than OS, this thesis will only consider PFS.

This new combination treatment has been submitted for marketing authorization approval to the European Medicines Agency (EMA) in 2019 and their application is currently under assessment [24]. The Food and Drug Administration (FDA) licensed this treatment in April 2020 [25] for adult patients with locally advanced or metastatic HER2-positive breast cancer, including patients with brain metastases[1], who have received one or more prior anti-HER2-based regimens in the metastatic setting.

[1]Brain metastases are tumors that have spread from the primary site of the cancer in the body, in this case the breast, to the brain.

The trial population for the PFS endpoint consists of 480 patients with HER2-positive cancer, some with presence of brain metastases. They were, via a randomization process, assigned to either the treatment arm (tucatinib plus trastuzumab and capecitabine) or the control arm (placebo plus trastuzumab and capecitabine). The assignment ratio was 2:1 meaning that 320 patients $(2/3)$ got the experimental treatment and 160 patients $(1/3)$ the standard treatment.

The first objective of the simulations is to investigate if alterations in the proportion of patients that are censored (censoring rates) induce bias in the estimates (median PFS time and HR). The second objective is to investigate the extent of the bias in the HR if the assumption of independent censoring is violated. If bias occur, the size and direction (under- or over-estimation) of the bias will be discussed.

In order to simulate censoring rates that are similar to those observed in the HER2CLIMB trial, the patients in the trial are divided into five subgroups, A-E, depending on whether they experienced a PFS event while on the study or discontinued the study before they experienced a PFS event. In table 3.1, these subgroups are presented. In the primary analysis, the investigators decided to censor all patients whose PFS event was not observed [1].

In the simulations, the statistical handling (censor or event) of patients who belongs to subgroups C, D and E, i.e. the patients who discontinued the study before experiencing a PFS event, changed to another anti-cancer therapy before experiencing a PFS event or who missed two or more assessments and experienced a PFS event in the following assessment, will be altered in order to see what bias may occur. In the HER2CLIMB trial, as in most clinical trials, the observations period is limited which causes administrative censoring. Subgroup A is the group of patients that were administratively censored when the observation period ended. However, in the simulations the observation period is until all patients either had an PFS event or were censored due to lost to follow-up or change of treatment which means that no administrative censoring is implemented in the simulations. This will not affect the bias, since administrative censoring is independent.

Subgroup	Description	No. patients	
		Control	Experimental
A	Patients on-going in the study who did not experience a PFS event at the time of database lock (admin. censoring).	9	54
B	Patients with observed PFS event.	97	178
C	Patients who discontinued the study before experiencing a PFS event.	2	7
D	Patients who changed to another anti-cancer therapy before experiencing a PFS event.	47	75
E	Patients who missed two or more assessments and experienced a PFS event in the following assessment.	5	6

Table 3.1: Description of patient subgroups, which are based on whether they experienced a PFS event while on the study or discontinued the study before they experienced a PFS event

The results from the study, concerning PFS, are presented in table 3.2.

HER2CLIMB study	Median PFS (95% C.I.)		HR (95% C.I.)	p
	Control	Experimental		
Estimate	5.6 (4.2, 7.1)	7.8 (7.5, 9.6)	0.54 (0.42, 0.71)	< 0.0001

Table 3.2: Estimated median PFS times for each treatment arm, and HR from the HER2CLIMB study that will be used to simulate similar survival data [1].

Chapter 4

Results and discussion

The simulations are divided into six cases, based on the subgroups listed in table 3.1. As mentioned, in the primary analysis for the HER2CLIMB trial, patients who did not experience a PFS event while being observed, or who missed two or more assessments and then experienced a PFS events were censored. The first case, case 0, replicates the results from the HER2CLIMB trial and is considered to be the base case. The five other cases are modifications of case 0, where the subgroups of patients that are censored are changed to be events in different combinations. This is done in order to assess the robustness of the published results and investigate potential bias caused by censoring. In table 4.1, all six cases are described and the censoring rates are presented. In short, the differences between the cases are how the subgroups C, D and E are treated in the statistical analysis. Either the patients in these subgroups (some or all of them) are censored or they are considered to have had their PFS event at the last observed PFS time, in both or one of the treatment arms. The censoring rate is the proportion of patients that are censored (not administratively) within each treatment arm. For instance, in case 0, 33.8% of the patients in the experimental treatment arm were censored.

In case 1, the patients who discontinued the study before experiencing a PFS event are considered to have had their PFS events, and in case 2, the patients who missed two or more assessments and experienced a PFS event in the following assessment are considered to have had their PFS event. These cases do not differ considerably from case 0 in terms of censoring rates. In case 3, non of the patients are censored, in neither of the arms, hence the censoring rates are equal to zero in both arms.

In case 4, the patients who were lost to follow-up or switched therapy are censored in the control treatment arm, but considered to have experienced an event in the experimental treatment arm. This is considered to be a conservative case, in the perspective of the regulator. The reason for this is that an observed PFS event decreases the PFS probability as compared to censoring. This will cause the estimated HR to be closer to 1 and, accordingly, the experimental treatment shows less difference in PFS compared to the control arm. On the contrary, in case 5, the opposite of case 4 is simulated, in the sense that the patients who were lost to follow-up or switched therapy are now censored in the experimental treatment arm but considered to have experienced an event in the control treatment arm. This is reasoned to be the ”best case scenario” from the perspective of the experimental treatment manufacturer, since the estimated HR will be as low as possible.

The first series of simulations, Simulation series 1, will implement independent censoring. This means that the censoring of patients will be distributed randomly within each treatment arm. The bias is expected to be small in case 0, 1 and 2 since independent censoring is a key assumption of the estimation methods and the censoring rates do not differ much between the cases. Also, the censoring proportion ratio among the treatment arms remain approximately the same. In case 3, where all of the previously censored patients are considered to be events, the estimated median PFS times should be shorter in both arms, since the probability of PFS is declined when there are more events, and therefore the median PFS is shortened. However, since the PFS event ratio between the treatment arms remain almost the same, the model should be able to handle this scenario well. On the other hand, in case 4 and 5, where the censoring/event assumption of the treatment arms contrast, the bias should be more significant.

In the second series of simulations, Simulation series 2, dependent and informative censoring is implemented in the experimental treatment arm. This is done by censoring the patients with a short ”true” PFS time to a higher extent than the other patients. More specifically, 80% of the censoring will be applied to the patients with a shorter PFS time than the median PFS time and, consequently, 20% of the censoring will be applied to the patients with equal or longer PFS time than the median PFS time. The censoring in the control arm is still independent, as in the Simulation series 1.

Simulation series 2 aims to recreate two frequently observed situations in oncology clini-

Case	Description	Censoring rate		No. PFS events	
		Control	Treatment	Control	Treatment
0	Patients who did not experience a PFS event while being observed (C + D), or who missed two or more assessments and then experienced a PFS events (E) are censored.	0.338	0.275	106	232
1	Patients who discontinued the study before experiencing a PFS event (C) are considered to have had a PFS events. The other subgroups are censored as in case 0.	0.325	0.253	108	239
2	Patients who missed two or more assessments and experienced a PFS event in the following assessment (D) are considered to have had a PFS event. The other subgroups are censored as in case 0.	0.306	0.256	111	238
3	Patients who did not experience a PFS event while being observed (C + D), or who missed two or more assessments and then experienced a PFS events (E) are considered to have had a PFS event.	0	0	160	320
4	The patients in the control treatment arm who did not experience a PFS event while being observed (C + D), or who missed two or more assessments and then experienced a PFS events (E) are censored (as case 0 but only for the control arm). The patients in the experimental treatment arm who did not experience a PFS event while being observed (C + D), or who missed two or more assessments and then experienced a PFS events (E) are considered to have had a PFS event (as case 3 but only for the experimental arm).	0.338	0	106	320
5	The patients in the control treatment arm who did not experience a PFS event while being observed (C + D), or who missed two or more assessments and then experienced a PFS events (E) are considered to have had a PFS event (as case 3 but only for the control arm). The patients in the experimental treatment arm who did not experience a PFS event while being observed (C + D), or who missed two or more assessments and then experienced a PFS events (E) are censored (as case 0 but only for the experimental arm). This is the reversed version of case 4.	0	0.275	160	232

Table 4.1: Description of six different censoring cases with the proportion of censored patients and number of censored patients in each treatment arm

cal trials. First, if the experimental treatment has an adverse toxicity profile, the result could be that the patients in this arm discontinue the treatment early since they do not tolerate the therapy, even if they have not experienced their PFS event yet. The second frequently observed situation is that patients switch to another anti-cancer therapy, despite that progression is not yet declared. This situation occurs in both treatment arms but can have different motivations behind, with varying consequences. In both cases, these patients are generally censored in the primary analysis. The issue with these situations is that these patients may not be well represented by the remaining patients in the study. This is because the discontinuation may be related to a worse health status or clinical progression, and therefore they may have a different progression or survival expectation than patients who remained in the study. An example of this, in the second situation, is when the physician is considering to change to another treatment for a patient but postpones as long as possible since the next treatment probably is less effective. These patients, for which the discontinuation is related to their health status, will be denoted as patients at *higher risk*. This means that the main assumption of independent censoring is violated and therefore it is interesting to simulate this scenario and investigate possible bias. In the first series of simulations, these patients decreased the survival probability, since the probability for them to experience their event were higher than for the rest of the patients. In this series of simulations, they are censored and thus the estimated HR is expected to be smaller than the true HR.

In [2], the author has conducted a similar simulation study where a comparison is made between a scenario where firstly, the patients who switch to another anti-cancer therapy in the control treatment arm are censored and secondly considered to have PFS events, whereas all patients in the experimental treatment arm have events. In other words, the author alters the censoring assumptions in the control arm, as compared to the simulation study in this thesis, where the alteration of censoring assumption is made for the experimental treatment arm. Carroll writes that the type of censoring of patients at higher risk that stop the control or experimental treatment before observed PFS event, should be avoided, since it leads to informative censoring.

In order to obtain a robust[1] result, 1000 simulations are made where in each the median survival, the hazard ratio and the log-rank statistic is computed. Then the median of these are computed as the point estimate and the 2.5'th and 97.5'th percentile are taken as confidence

[1]The results vary for each simulation run due to variability and 1000 simulation repetitions gives robustness of the estimates.

bounds, using the bootstrap method [26] and presented in tables. Furthermore, Kaplan-Meier estimated survival curves are plotted and presented in figures to illustrate the results. A more detailed technical description of the simulations is presented in Appendix A.

4.1 Simulation series 1 - Independent censoring

The estimated median PFS times and hazard ratios are presented in figures 4.1 and 4.2. They are essentially equal for case 0, 1 and 2 with negligible HR bias in case 2. In addition, the HR for case 3 is similar to case 0, with negligible HR bias. In the cases where the previously censored patients are treated as events in the analysis, the median PFS times are reduced, see for instance case 3 with median PFS 4.12 and 6.16 in the control and experimental treatment arm respectively. Furthermore, in case 4 and 5, the HR bias is quite large. In case 4, the HR is estimated to be 0.803 which means that the efficacy of the experimental treatment is estimated to be much smaller than in the HER2CLIMB study. On the other hand, in case 5, the HR is estimated to be 0.370, which indicates that the experimental treatment efficacy is better than in the HER2CLIMB study. The reason for the decreased median PFS times is that in these cases there are more PFS events that lower the PFS probability. In case 3, the HR is still rather unbiased which is explained by the fact that the previously censored patients are treated as event, in both treatment arms, as opposed to case 4 and 5, where they are treated differently depending on which arm they belong to. This means that, in case 4, the hazard is lower for the control treatment arm but higher for the experimental treatment arm, resulting in a greater HR. For case 5, it is the other way around. This explain why the HR is biased in case 4 and 5, either towards 1 (in case 4), meaning that the efficacy is worse, or towards 0 (in case 5), meaning that the efficacy is better.

The HR bias presented in figure 4.2 is not due to violation of the independent censoring assumption, but rather the fact that they are considered to have their PFS event earlier than they would had been if they were censored. Another reason for the bias in case 4 and 5 is that the censoring proportion ratios among the treatment arms differ significantly. For example, if one arm has a much larger censoring rate than the other, the bias is in favour of the treatment arm with a higher censoring rate.

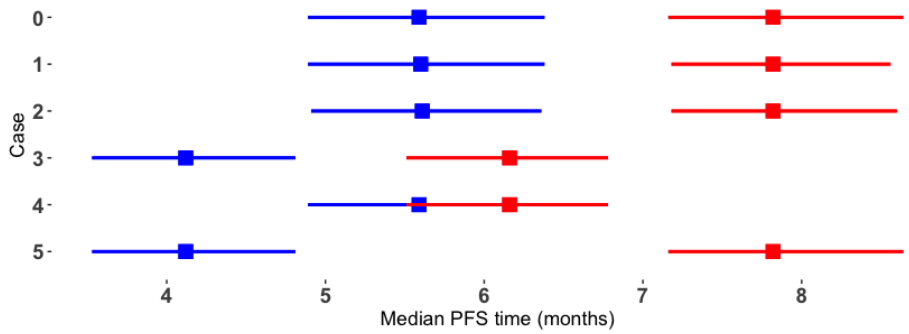
The width of the confidence intervals varies and is the narrowest for case 3, and widest for case 4. Thus, in these cases, precision of the estimates is best and worse respectively, as

compared to the other cases. An explanation for this is that cases that involves more censoring, as case 3, induces uncertainty into the Cox model, since an exact event time is not known, only a minimum. If independent censoring can be assumed, case 3 is the best alternative from a regulatory point of view, since no significant bias is induced and the precision is the best.

An HR of 0.54 implies a rather large difference in treatment efficacy (PFS) between the treatment arms. In order to assess whether there is a difference in PFS for patients treated with the experimental treatment compared to those who received the standard treatment, it is important that the upper 95% confidence bound of the HR is below 1. In this case, the true simulated HR is 0.54, which is relatively far away from 1. This reduces the importance of the confidence intervals, since the experimental treatment still show a statistically significant effect, even if we consider the upper bound of the interval. However, if the HR would be higher, which is the case in many oncology studies [2], the confidence intervals could possibly cover 1. This would imply that the experimental treatment does not have a statistically lower PFS than the control arm. A simulation series with HR$= 0.8$ and independent censoring was made (not presented) where the 95% confidence intervals covered 1, meaning that there are no statistically significant difference in the efficacy of the treatments, on a 5% significance level.

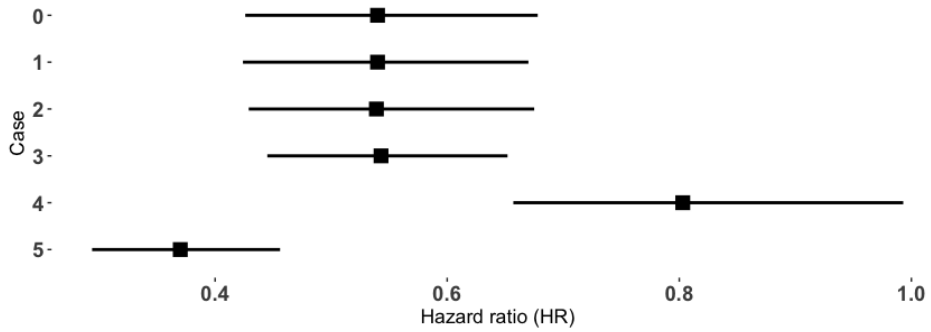
In all cases except case 4, the log-rank statistic supports that the survival curves for each treatment are not statistically equal.

[2] In [27] a literature study was done of non-small cell lung cancer trials reported between January 1991 and November 2010. In figure 2A, HRs are plotted and a major part is in the interval $0.6 - 1$. In [28], oncology trials in PubMed (a database consisting of biomedical literature, life science journals, and online books) were investigated in order to study the correlation between compare PFS to OS as endpoints. In table 2, several HRs around 0.8 are presented.



Case	Median PFS (95% C.I.) Control	Median PFS (95% C.I.) Experimental
0	5.59 (4.89, 6.38)	7.82 (7.16, 8.64)
1	5.60 (4.89, 6.38)	7.82 (7.18, 8.56)
2	5.61 (4.91, 6.36)	7.82 (7.18, 8.60)
3	4.12 (3.53, 4.81)	6.16 (5.51, 6.78)
4	5.59 (4.89, 6.38)	6.16 (5.51, 6.78)
5	4.12 (3.53, 4.81)	7.82 (7.16, 8.64)

Figure 4.1: The estimated median PFS with 95% confidence intervals for the control treatment group and experimental group respectively, from Simulation series 1.



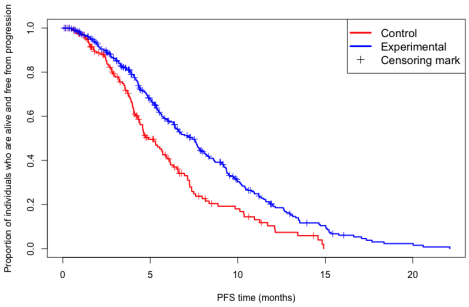
Case	HR (95% C.I.)	HR bias (HR_est - HR_true)	Log-rank statistic
0	0.540 (0.426, 0.678)	0	25.8
1	0.540 (0.424, 0.670)	0	26.8
2	0.539 (0.429, 0.675)	-0.001	27.5
3	0.543 (0.445, 0.652)	0.003	38.0
4	0.803 (0.657, 0.993)	0.263	3.60
5	0.370 (0.294, 0.456)	-0.170	91.5

Figure 4.2: The estimated HR with 95% confidence intervals, HR bias and log-rank statistics for the control treatment group and experimental group respectively, from Simulation series 1.

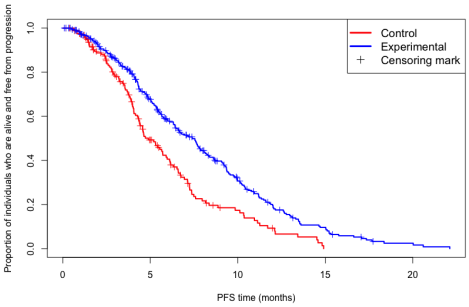
In figure 4.3, the estimated Kaplan-Meier curves are plotted. The survival curves of case 0-3 are quite similar, which support the similarity of the estimated HR for these cases. A visible difference is that, when there is censoring involved, as in case 0-2, the curves are slightly higher, as opposed to for instance in case 3, where patients with no observed progression are included as events, in both treatment arms. This is also expected since the censoring do not decrease the survival probability, as an observed event does. However, the difference is so small, on average, so it is not shown in the hazard ratio estimates in figure 4.2, since both survival curves are lowered.

In case 4, the high HR of 0.803 is reflected in the estimated survival curves, see figure 4.3e. They are barely separable which implies that the difference in PFS experience between the treatment arm is quite small. This might be explained by the fact that the patients that are lost to follow-up or with therapy are censored in the control group, resulting in a higher survival curve, whereas these patients are considered to be events in the experimental treatment group, which decreases the survival curve. In case 5 on the other hand, the curves are even more separated than in case 0, 1, 2 and 3 which also is a support for the low HR.

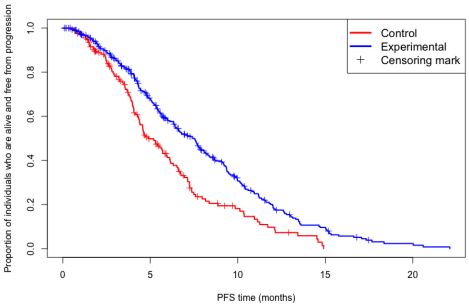
The Kaplan-Meier survival curves and the HR are not based on the same models, but they show the same tendency of the efficacy and survival experience between the treatment arms.



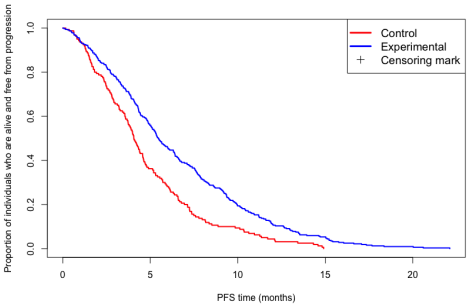
(a) Case 0, where the patients who are lost to follow-up or switched therapy are censored, in both treatment arms.



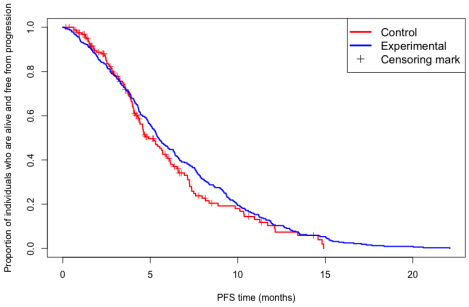
(b) Case 1, where the patients who discontinued the study before experiencing a PFS event are considered to have had an PFS events, in both treatment arms. The other subgroups are censored as in case 0.



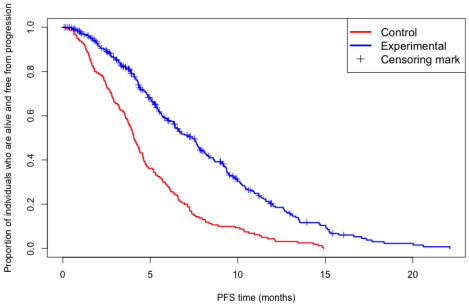
(c) Case 2, where the patients who missed two or more assessments and experienced a PFS event in the following assessment are considered to have had an observed PFS event, in both treatment arms. The other subgroups are censored as in case 0.



(d) Case 3, where the patients who are lost to follow-up or switched therapy are considered to have an observed PFS event, in both treatment arms.



(e) Case 4, where the patients in the control treatment arm who were lost to follow-up or switched therapy are censored, whereas these patients in the experimental arm are considered to have an observed PFS event.



(f) Case 5, where the patients in the experimental arm treatment arm who are lost to follow-up or switched therapy are censored, whereas these patients in the control arm are considered to have an observed PFS event.

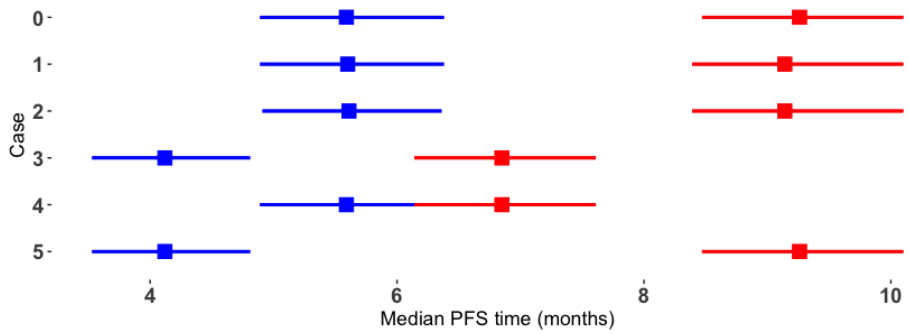
Figure 4.3: Kaplan-Meier survival curves for all cases with independent censoring, from Simulation series 1.

4.2 Simulation series 2 - Dependent censoring

Firstly, in case 0, 1, 2 and 3, where the patients in the experimental treatment arm who were lost to follow-up or switched therapy were censored, the median PFS times are prolonged (≈ 9.2) compared to ≈ 7.8 in Simulation series 1, see figure 4.4. Also, the HR is underestimated (≈ 0.42), which implies that the treatment efficacy, in terms of PFS in the experimental treatment arm, is overestimated, see figure 4.5. Note that the estimated HR for case 3, where the patients who did not switch to other anti-cancer therapies or discontinued the study before progression were considered events, is also underestimated but not as much as in case 0, 1 and 2.

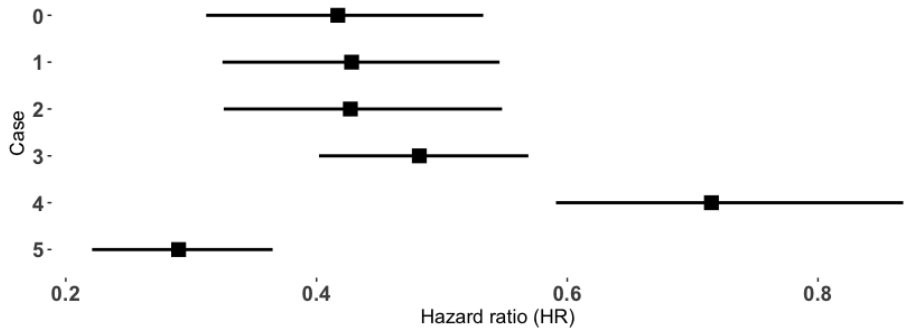
The median PFS times in case 0, 1, 2 and 5 in the experimental treatment are longer since the patients who earlier shortened the overall median survival now are censored, which implies that their PFS event is not observed. This is also the reason for the estimated HR to be smaller than the true HR. If this type of censoring would occur in a trial, which is expected at least partly, the HR would be underestimated. This kind of censoring seems likely, for instance when the experimental treatment show a high toxic profile. In those cases, in order to avoid underestimating the HR in favor for the experimental treatment, it would be better to consider a case where the worst estimate of HR, in terms of experimental treatment efficacy, is tested (case 4), at least as a supplementary analysis. The confidence interval for the HR for case 4, where the patients who are lost to follow-up or switch treatment are censored in the control arm but considered as PFS events in the experimental arm, are wider than the others. In this example it still shows a statistically significant treatment effect, since the hazard ratio is low. However, if HR was higher, around 0.8, it would be anticipated that the confidence intervals include 1.

The log-rank statistics for all cases except case 4 imply that the null hypothesis, which is that the survival curves are equivalent for the two treatment groups, can be rejected since they all are greater than $\chi^2(1) = 3.74$ on a 5% significance level. However, this test is only valid under the assumption of independent censoring, which is not met in this series of simulations.



Case	Median PFS (95% C.I.) Control	Median PFS (95% C.I.) Experimental
0	5.59 (4.89, 6.38)	9.26 (8.47, 10.1)
1	5.60 (4.89, 6.38)	9.14 (8.39, 10.1)
2	5.61 (4.91, 6.36)	9.14 (8.39, 10.1)
3	4.12 (3.53, 4.81)	6.85 (6.14, 7.61)
4	5.59 (4.89, 6.38)	6.85 (6.14, 7.61)
5	4.12 (3.53, 4.81)	9.26 (8.47, 10.1)

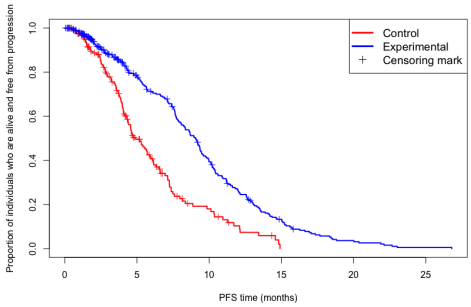
Figure 4.4: The estimated median PFS with 95% confidence intervals for the control treatment group and experimental group respectively, from Simulation series 2.



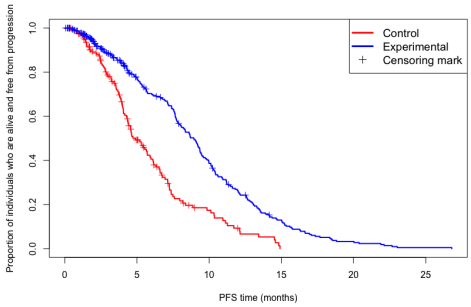
Case	HR (95% C.I.)	HR bias (HR_est - HR_true)	Log-rank statistic
0	0.417 (0.312, 0.533)	-0.123	50.6
1	0.428 (0.325, 0.546)	-0.112	49.0
2	0.427 (0.326, 0.548)	-0.113	50.1
3	0.482 (0.402, 0.569)	-0.058	52.0
4	0.715 (0.591, 0.868)	0.175	8.14
5	0.290 (0.221, 0.365)	-0.250	135

Figure 4.5: The estimated HR with 95% confidence intervals, HR bias and log-rank statistics for the control treatment group and experimental group respectively, from Simulation series 2.

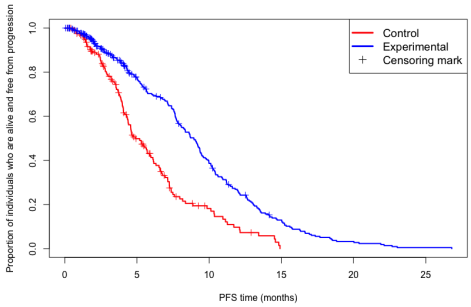
In figure 4.6, the estimated Kaplan-Meier survival curves are plotted, for each case. In case 0, 1, 2 and 5, where the patients who are lost to follow-up or switched treatment are censored in the experimental treatment arm, larger differences between the experimental and control treatment arms are seen, which confirms the lower estimated hazard ratios.



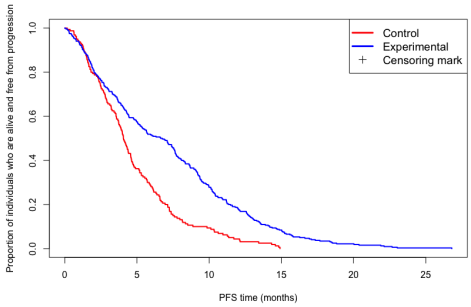
(a) Case 0, where the patients who are lost to follow-up or switched therapy are censored, in both treatment arms.



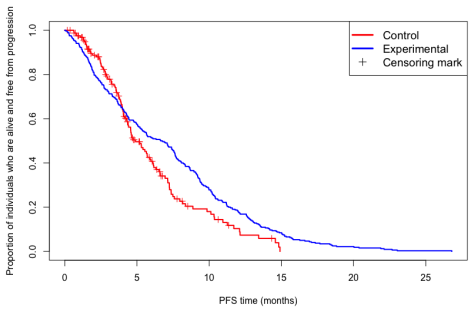
(b) Case 1, where the patients who discontinued the study before experiencing a PFS event are considered to have had an PFS events, in both treatment arms. The other subgroups are censored as in case 0.



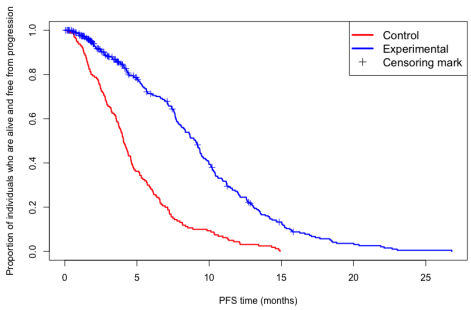
(c) Case 2, where the patients who missed two or more assessments and experienced a PFS event in the following assessment are considered to have had an observed PFS event, in both treatment arms. The other subgroups are censored as in case 0.



(d) Case 3, where the patients who are lost to follow-up or switched therapy are considered to have an observed PFS event, in both treatment arms.



(e) Case 4, where the patients in the control treatment arm who were lost to follow-up or switched therapy are censored, whereas these patients in the experimental arm are considered to have an observed PFS event.



(f) Case 5, where the patients in the experimental arm treatment arm who are lost to follow-up or switched therapy are censored, whereas these patients in the control arm are considered to have an observed PFS event.

Figure 4.6: Kaplan-Meier survival curves for all cases with dependent censoring, from Simulation series 2 with dependent censoring.

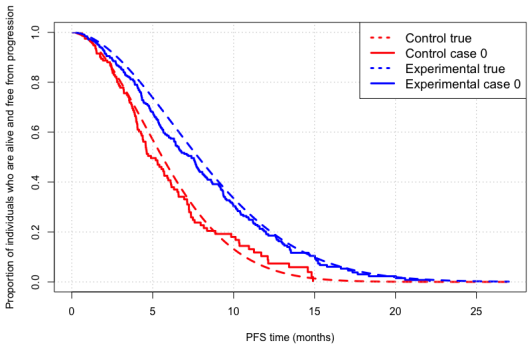
4.3 Comparison of the simulations

From a regulatory perspective (for instance the authorities responsible for granting marketing authorizations or approving reimbursement of treatments' cost), it is interesting to compare case 0, that replicates the HER2CLIMB results, and case 4, where all patients that are lost to follow-up or switched therapy are censored in the control arm but considered to have had the event in the experimental arm. As mentioned previously, case 4 is the most conservative case. This is because the events in the experimental arm decreases the survival probability for this arm but not the control arm, where these patients are censored. Thus, the hazard ratio increases and the experimental treatment performs worse. In figure 4.7, the Kaplan-Meier survival curves are plotted together with the true survival curve, for each case and simulation series respectively. In figure 4.7a, where case 0 and independent censoring is depicted, the survival curves are quite similar to the true simulated curve. This is expected since the independent censoring assumption of the Kaplan-Meier estimation method is met. On the other hand, in figure 4.7b, where case 0 and independent censoring is depicted, the experimental treatment curve is significantly underestimated compared to the true simulated curve. This is because, as explained previously, more events lower the estimated PFS curve.

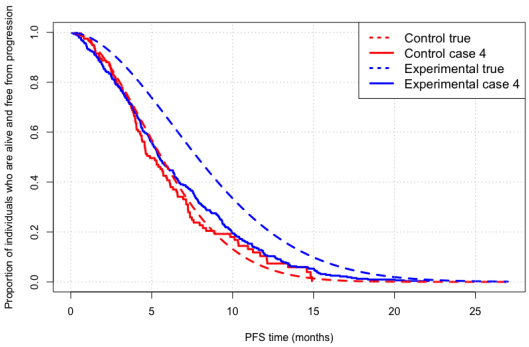
In figure 4.7c, where case 0 and dependent censoring is depicted, the experimental treatment curve is overestimated compared to the true curve, meaning that the free-of-progression or survival probability is estimated to be better than it actually is for the experimental patient group. However, in figure 4.7d, the same curve is underestimated. This confirms that case 4 provides a more conservative choice in favour of the control treatment, but also that case 0 imply that the experimental treatment efficacy is overestimated, if the censoring is dependent.

If the independent censoring assumption is believed to be met, at least to an acceptable extent with negligible proportions of patients where dependent censoring is suspected, case 4 is quite extreme and may be overly conservative. Then case 3 provides a better alternative, as previously mentioned in 4.1. Nevertheless, case 4 provides an estimate of the minimum efficacy of the experimental treatment. However, if dependent censoring is suspected, case 4 provides a more justifiable conservative scenario that can serve as guidance to regulatory services.

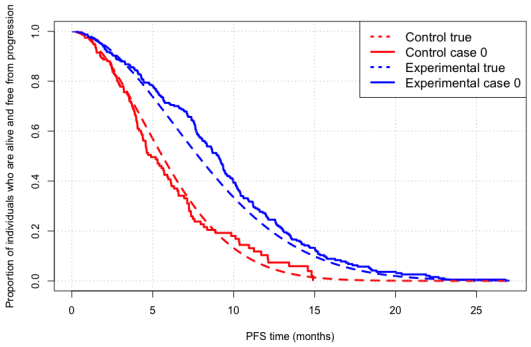
The relation of the censoring time to the event time is crucial, since the estimation methods depend on the assumption of independent censoring. In many cases, such as administrative



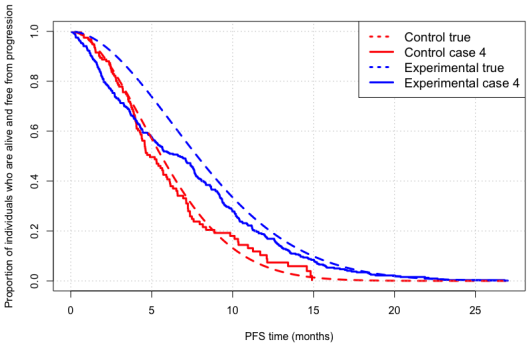
(a) K-M survival curves for case 0 with independent censoring together with the true survival curve.



(b) K-M survival curves for case 4 with independent censoring together with the true survival curve.



(c) K-M survival curves for case 0 with dependent censoring together with the true survival curve.



(d) K-M survival curves for case 4 with dependent censoring together with the true survival curve.

Figure 4.7: Kaplan-Meier survival curves for comparison of case 0, when patients that are lost to follow-up or switch therapy are censored, and case 4, where those patients are considered to have a PFS event, for independent and dependent censoring simulations respectively, together with the true survival curves, if HR=0.54.

censoring or loss to follow-up, independent censoring can be assumed without any resulting estimation bias. If this is the case, it is unnecessarily conservative to consider a scenario such as case 4. However, in some cases, such as when the experimental treatment has a high toxicity profile depending on the overall health status, the censoring time can be related to the PFS time. Therefore, it would be useful to know why the patients did not stay in the trial or switched to another anti-cancer therapy. If it is related to their survival, for example if the disease status has gotten worse, the HR will be underestimated, meaning that the treatment efficacy will be overestimated. Given that the censoring mechanisms are unknown, it is therefore of uppermost importance that the primary analysis are followed by supplementary analysis which simulates a range of censoring schemes, in order to evaluate the robustness of the primary analysis. Based on for instance the toxicity profile of the experimental treatment, a customized supplementary analysis may be done, in order to draw confident and decisive conclusions about the efficacy of the experimental treatment. In the HER2CLIMB study the censoring rates were approximately around 30% in both treatment arms which is a rather large proportion. It is reasonable to believe that at least a part of the censoring is related to the patients' PFS. Therefore, it is important to investigate the underlying reason why the patients are censored.

Another important aspect is to investigate the censoring proportions. It is noted that the bias in case 4 is in favour of the experimental treatment arm. This is due to the fact that this arm has the greatest share of censoring. In other words, the bias punishes the arm that has the highest censoring rate. In case 5, where the role of censoring/events is the opposite, the contrary is true, i.e. the bias is in favour the control treatment arm. Therefore, it is also important to investigate the censoring proportions in the treatment arms, in order to choose the most conservative case.

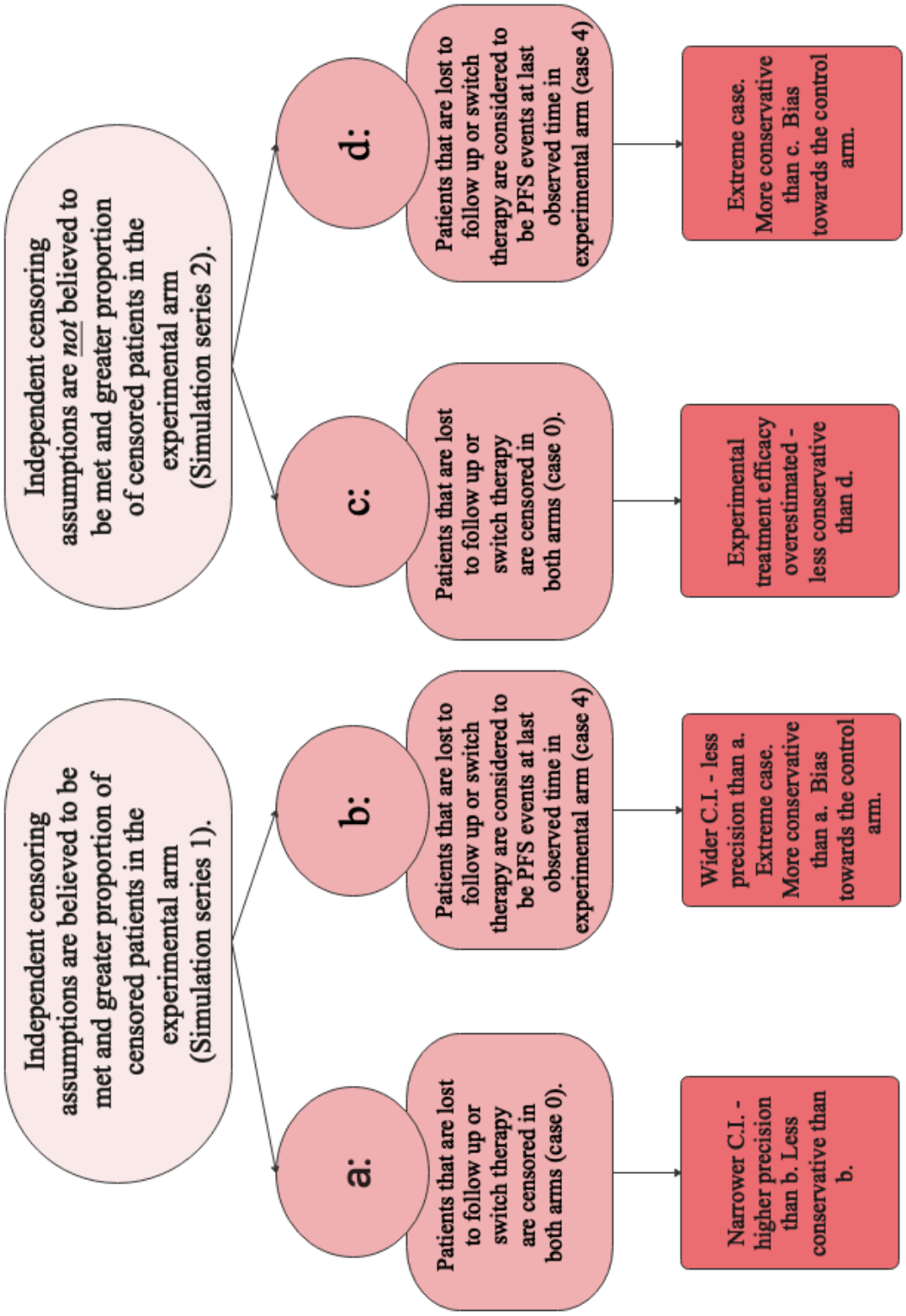


Figure 4.8: Illustration of the comparison between case 0 and case 4 in terms of precision and conservativeness, for Simulation series 1 and 2 respectively.

Chapter 5

Conclusions

In oncology clinical trials, where an experimental treatment is compared to a control treatment, progression-free survival (PFS) is a common primary endpoint. If a patient discontinue before progression of the disease or death, the PFS time is unobserved and hence, an assumption about the patient's PFS time need to be made. This is called censoring and induces uncertainty to the estimates. In some cases, it is reasonable to believe that the censored patient is well represented by the remaining patients in the trial which means that independent censoring is a valid assumption. However, in some cases patients discontinue early due to intolerability of the experimental treatment. In these cases, the censoring times are probably related to the PFS times which means that the independent censoring assumption is violated. This can give biased estimates since independent censoring is a key assumption of the statistical estimation methods such as Kaplan-Meier survival curves and the Cox model.

The objective of this thesis was to investigate the possible bias due to censoring of patients who discontinue early in a phase 3 oncology clinical trial. An estimate of the possible bias was computed by simulating survival data that was based on a recent oncology clinical trial, the HER2CLIMB study. Patient subgroups were identified based on whether the patients discontinued early and, in that case, why. Based on these subgroups, six different cases were simulated and analysed. Furthermore, the cases were investigated under two different censoring assumptions, independent (Simulation series 1) and dependent censoring (Simulation series 2).

The HER2CLIMB trial resulted in a low HR (0.54) which makes the results robust to different censoring scenarios. A higher HR would be more sensitive to changes of the censoring

rates. When independent censoring was implemented, the HR was unbiased in the cases where the censoring rates are similar to what they were in the HER2CLIMB study. It was also unbiased when all patients are considered to have had their event at the last observed time. However, in a more conservative case the HR is biased. The direction of the bias depends on the censoring proportion in each arm.

When dependent censoring was implemented, the cases that were similar to the HER2CLIMB study in terms of censoring rates, resulted in an underestimated HR. In a more conservative setting, the HR was overestimated but not as much as in the independent censoring setting. Therefore, if the reason for the patient to discontinue could be related to their time to progression or death in some way, the assumption of independent censoring is violated and supplementary analyses are encouraged. Moreover, it is important to investigate the censoring proportions in each arm. The conservative choice depends on which arm has the greater share of censoring.

To summarise, censoring of patients often induces uncertainty into the model since the independent censoring assumption may or may not be met. A way of reducing this uncertainty is to investigate why the patients are lost to follow-up and consider a more specific censoring scheme, based on the reason for discontinuation.

5.1 Further research

Firstly, this simulation study was following a specific oncology clinical which makes the result specific for this study. However, the results should be applicable to other oncology studies with similar HR, censoring rates and PFS patterns. In a larger scope project, it would have been interesting to consider a third series of simulation where an hazard ratio of 0.8 in combination with dependent censoring would be investigated in terms of censoring bias.

Furthermore, the implementation of dependent censoring can be varied, for instance by investigating the censoring bias if patients at lower risk are censored to an higher extent than the other patients. This would probably give less bias than in Simulation series 2, since censoring later on in the follow-up period do not influence the efficacy estimate as much as censoring earlier on.

In order to obtain more practically accurate results, a more complicated survival times model could have been implemented, such as a mixture model that is suggested in [22]. Also, other independent variables, such as a categorical variable for health status (in HER2CLIMB study they use the Eastern Cooperative Oncology Group (ECOG) performance status) could have been considered for a stratified Cox model as well as weighted censoring.

In the dependent censoring simulation, the percentage of censoring in low and high respectively is not exactly 80% an differs slightly for each case. This should not lead to that much difference since first, 80% is chosen just to show an example and the important point is that it is more censoring for those at higher risk, and second, it is small differences. Nevertheless, it would have been desirable to further improve this part of the code.

The log-rank statistics that are presented in figure 4.2 and 4.5 is the 50'th percentile log-rank statistics of the 1000 simulation run for each simulation series respectively. It is not completely investigated if this is a valid representation of the statistic. In a larger scope project, this would have been investigated further.

Appendix A

Technical description of the simulations

The simulations are done in R, a software environment for statistical computations. A simulated dataset consists of four column vectors with data about the patient' id, observed PFS time, treatment arm and censoring status. All vectors are of length length 480, which is the number of patients in the HER2CLIMB study [1].

The treatment status vector contains 160 0's (control treatment arm) and 320 1's (experimental treatment arm).

The PFS event time variable E is simulated from a Weibull distribution using the inverse cumulative hazard functions that is described in [21]. A standard uniform random variable U_E is drawn and the Weibull survival times are simulated by

$$E = \left(-\frac{\log(U)}{\lambda \exp(\beta \cdot \texttt{trt|})} \right)^{1/\rho}$$

where ρ is the scale parameter, λ sis the shape parameter, β is the treatment status variable coefficient. The censoring time variable C is simulated from an exponential distribution with intensity parameter ω. The inverse cumulative hazard functions is then

$$C = -\frac{\log(U)}{\omega \exp(\beta \cdot \texttt{trt|})}$$

where U_C is a standard uniform random variable. The observed PFS time, T, is then computed using equation 2.1.

The median PFS time, t_{median}, is set to be 5.6, which is the estimated median PFS for the con-

trol group in the HER2CLIMB study. The shape parameter λ is then given by equation 2.2. Secondly, the scale parameter ρ is fitted such that the median PFS for the experimental treatment group becomes approximately 7.8. Choosing it as $\rho = 1.85$ is shown to be appropriate. These parameter values will be kept fixed for all cases and simulation series in order to obtain comparable results.

In order to obtain the wanted censoring rates for the different censoring cases and simulation series, ω is fitted for each case and censoring rate.

In the first simulation series, where independent censoring is implemented, one parameter value is fitted for each case, see table A.1.

Case	Intensity parameter ω	
	Control	Experimental
0	0.0728	0.0738
1	0.0690	0.0668
2	0.0640	0.0678
3	-	-
4	0.0728	-
5	-	0.0738

Table A.1: The intensity parameter ω for the censoring random variable, for each case in Simulations series 1, with independent censoring.

In the second simulation series, where dependent censoring is implemented, two parameter values are fitted since the censoring rate is higher for the patients with $E < t_{median}$ and lower for $E \geq t_{median}$, see table A.2.

Case	Intensity parameter ω		
	Control	Experimental	
		Low (%)	High (%)
0	0.0728	0.239 (79.9)	0.0178 (20.1)
1	0.0690	0.214 (80.2)	0.016 (19.8)
2	0.0640	0.214 (80.2)	0.016 (19.8)
3	-	-	-
4	0.0728	-	-
5	-	0.239 (79.9)	0.0178 (20.1)

Table A.2: The intensity parameter ω for the censoring random variable, for each case in Simulation series 2. Low denotes the category for the censoring of patients with event times that are shorter than the median PFS time. High denotes the category for the censoring of patients with event times that are equal to or longer than the median PFS time.

The median PFS times, Kaplan-Meier survival curves, the log rank statistics and the Cox model are estimated using the `survival` package in R, using the `survfit`, `survdiff` and `coxph` functions. These functions are applied on a formula with a `Surv` object on the left hand side and the `trt` variable on the right hand side.

Given a dataset with observed PFS times `time`, censoring status `delta` and treatment status `trt`, the following R code computes the Kaplan-Meier survival curves, the log-rank statistic and the Cox proportional hazards model.

```
KM.model <- survfit(Surv(time, delta) ~ trt, dataset)

log.rank <- survdiff(Surv(time, delta) ~ trt, dataset)

cox.model <- coxph(Surv(time, delta) ~ trt, dataset)
```

Bibliography

[1] Rashmi K. Murthy et al. “Tucatinib, Trastuzumab, and Capecitabine for HER2-Positive Metastatic Breast Cancer”. In: *New England Journal of Medicine* 382.7 (2019), pp. 597–609. DOI: 10.1056/NEJMoa1914609. URL: https://www.nejm.org/doi/full/10.1056/NEJMoa1914609.

[2] Kevin J. Carroll. “Analysis of progression-free survival in oncology trials: Some common statistical issues”. In: *Pharmaceutical Statistics* 6.2 (2007), pp. 99–113. ISSN: 1539-1604. DOI: 10.1002/pst.251. URL: %3CGo%20to%20ISI%3E://WOS:000247218900004.

[3] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text, Third Edition*. Statistics for Biology and Health. New York, NY: Springer New York, 2012. URL: http://ludwig.lub.lu.se/login?url=https://link.springer.com/10.1007/978-1-4419-6646-9%20http://dx.doi.org/10.1007/978-1-4419-6646-9.

[4] F. Campigotto and E. Weller. “Impact of informative censoring on the Kaplan-Meier estimate of progression-free survival in phase II clinical trials”. In: *J Clin Oncol* 32.27 (2014), pp. 3068–74. ISSN: 1527-7755 (Electronic) 0732-183X (Linking). DOI: 10.1200/JCO.2014.55.6340. URL: https://www.ncbi.nlm.nih.gov/pubmed/25113767.

[5] V. Prasad and U. Bilal. “The role of censoring on progression free survival: oncologist discretion advised”. In: *Eur J Cancer* 51.16 (2015), pp. 2269–71. ISSN: 1879-0852 (Electronic) 0959-8049 (Linking). DOI: 10.1016/j.ejca.2015.07.005. URL: https://www.ncbi.nlm.nih.gov/pubmed/26259493.

[6] Simon Fink. “Sensitivity Analyses for Informative Censoring in Time-to-Event Clinical Trials”. Thesis. 2015.

[7] Web Page. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/human-epidermal-growth-factor-receptor-2.

[8] Lawrence M. Friedman et al. *Fundamentals of clinical trials*. Fifth edition. Cham: Springer, 2015, xxi, 550 pages. ISBN: 9783319185385 (alk. paper) 3319185381 (alk. paper).

[9] Web Page. 2018. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/.

[10] US Food and Drug Administration. "Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics". In: (2018). URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics.

[11] European Medicines Agency. *ICH E9 statistical principles for clinical trials*. Report. 1998. URL: https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials.

[12] Elizabeth A Eisenhauer et al. "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)". In: *European journal of cancer* 45.2 (2009), pp. 228–247.

[13] Lothar R. Pilz, Christian Manegold, and Gerald Schmid-Bindert. "Statistical considerations and endpoints for clinical lung cancer studies: Can progression free survival (PFS) substitute overall survival (OS) as a valid endpoint in clinical trials for advanced non-small-cell lung cancer?" In: *Translational lung cancer research* 1.1 (2012), pp. 26–35. ISSN: 2218-6751 2226-4477. DOI: 10.3978/j.issn.2218-6751.2011.12.08. URL: https://pubmed.ncbi.nlm.nih.gov/25806152%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4367593/.

[14] E. D. Saad and M. Buyse. "Statistical controversies in clinical research: end points other than overall survival are vital for regulatory approval of anticancer agents". In: *Ann Oncol* 27.3 (2016), pp. 373–8. ISSN: 1569-8041 (Electronic) 0923-7534 (Linking). DOI: 10.1093/annonc/mdv562. URL: https://www.ncbi.nlm.nih.gov/pubmed/26578738.

[15] C. Davis et al. “Availability of evidence of benefits on overall survival and quality of life of cancer drugs approved by European Medicines Agency: retrospective cohort study of drug approvals 2009-13”. In: *BMJ* 359 (2017), j4530. ISSN: 1756-1833 (Electronic) 0959-8138 (Linking). DOI: 10.1136/bmj.j4530. URL: https://www.ncbi.nlm.nih.gov/pubmed/28978555.

[16] Dirk F. Moore. *Applied Survival Analysis Using R*. Use R! Cham: Springer International Publishing, Imprint: Springer, 2016. ISBN: 9783319312453. URL: http://dx.doi.org/10.1007/978-3-319-31245-3%20http://ludwig.lub.lu.se/login?url=https://link.springer.com/10.1007/978-3-319-31245-3.

[17] David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied survival analysis : regression modeling of time-to-event data*. 2nd. Wiley series in probability and statistics. Hoboken, N.J.: Wiley-Interscience, 2008, xiii, 392 p. ISBN: 9780471754992 (cloth) 0471754994 (cloth).

[18] Tamás Rudas. *Handbook of probability : theory and applications*. Los Angeles: Sage Publications, 2008, xv, 469 p. ISBN: 9781412927147 (cloth alk. paper). URL: Table%20of%20contents%20only%20http://www.loc.gov/catdir/toc/ecip0726/2007036434.html.

[19] D.R. Cox and D. Oakes. *Analysis of survival data*. Jan. 2018, pp. 1–201. DOI: 10.1201/9781315137438.

[20] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*. John Wiley & Sons, 1991.

[21] Ralf Bender, Thomas Augustin, and Maria Blettner. “Generating survival times to simulate Cox proportional hazards models”. In: *Statistics in medicine* 24.11 (2005), pp. 1713–1723.

[22] M. J. Crowther and P. C. Lambert. “Simulating biologically plausible complex survival data”. In: *Stat Med* 32.23 (2013), pp. 4118–34. ISSN: 1097-0258 (Electronic) 0277-6715 (Linking). DOI: 10.1002/sim.5823. URL: https://www.ncbi.nlm.nih.gov/pubmed/23613458.

[23] D.R. Cox and D. Oakes. *Analysis of Survival Data*. Taylor Francis, 1984. ISBN: 9780412244902. URL: https://books.google.se/books?id=Y4pdM2soP4IC.

[24] Press Release. 2020. URL: https://investor.seattlegenetics.com/press-releases/news-details/2020/EMA-Validates-Seattle-Genetics-Marketing-Authorization-Application-for-Tucatinib-for-Patients-with-Locally-Advanced-or-Metastatic-HER2-Positive-Breast-Cancer/default.aspx.

[25] Press Release. 2020. URL: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-tucatinib-patients-her2-positive-metastatic-breast-cancer.

[26] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1997. ISBN: 9780521574716. DOI: DOI:10.1017/CBO9780511802843. URL: https://www.cambridge.org/core/books/bootstrap-methods-and-their-application/ED2FD043579F27952363566DC09CBD6A.

[27] K. Hotta et al. "Progression-free survival and overall survival in phase III trials of molecular-targeted agents in advanced non-small-cell lung cancer". In: *Lung Cancer* 79.1 (2013), pp. 20–6. ISSN: 1872-8332 (Electronic) 0169-5002 (Linking). DOI: 10.1016/j.lungcan.2012.10.007. URL: https://www.ncbi.nlm.nih.gov/pubmed/23164554.

[28] C Wayant and M Vassar. "A comparison of matched interim analysis publications and final analysis publications in oncology clinical trials". In: *Annals of Oncology* 29.12 (2018), pp. 2384–2390.