UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

# Comparison of duration of response endpoint in oncology

Author: **Thuy Linh Do**
Supervisors: **Alexandre Lambert, Catherine Legrand**
Reader: **Ingrid Van Keilegom**
Academic year 2022–2023
Master [120] en statistique, orientation biostatistiques

## Preface.

This dissertation has been prepared in partial fulfillment of the requirements for the master degree in statistics delivered by the Universite Catholique de Louvain.

## Acknowledgment.

I would like to thank all the people who helped me, in one way or another, to complete this master thesis.

My first thanks go to my two promoters, Alexandre Lambert and Catherine Legrand. Thank you for your availability, advice and support throughout the production of this master thesis. Sharing your experience and knowledge helped me to better understand the various difficulties encountered during the course of this work, and above all to learn a great deal about the world of clinical trials and oncology.

Secondly, I would like to thank the entire Qubes team for integrating me into their team enriching these last few months with daily exchanges, sound advice and interesting sharing.

Finally, I would like to thank my family and friends for their advice and opinions at every stage of my work. Their unfailing support was indispensable for me to complete this master thesis.

# Acronyms

**AUC** Area Under Curve. 15, 30

**CI** Confidence Interval. 20, 22, 36, 40

**CR** Complete Response. 2

**CRR** Cumulative Response Rate. 10, 20

**DOR** Duration Of Response. 3, 9, 10, 19, 20, 22, 27, 29, 36, 40, 43, 44

**EMA** European Medicines Agency. 1, 10

**FDA** Food and Drug Administration. 1, 35

**HPV** Human Papilloma Virus. 35

**HR** Hazards Ratio. 8–10, 19, 20, 22, 23, 29, 36, 40, 43

**IQR** Interquartile Range. 20

**KM** Kaplan-Meier. 9, 10, 15, 22, 44

**MDOR** Mean Duration Of Response. 22, 40

**OR** Odds Ratio. 19, 20, 29, 36

**ORR** Objective Response Rate. 1, 35

**OS** Overall Survival. 1, 2, 35, 36

**PBIR** Probability of Being In Response. 3, 9–12, 14, 19, 20, 22, 23, 27, 30, 40, 43, 44

**PD** Progression Disease. 2

**PFS** Progression-Free Survival. 1, 2, 6, 11, 18, 35, 36

**PR** Partial Response. 2

**RECIST** Response Evaluation Criteria In Solid Tumors. 2

**SCC** Oropharyngeal Squamous Cell Carcinoma. 35

**SCCHN** Squamous Cell Carcinoma of Head and Neck. 35

**SD** Stable Disease. 2

**TIR** Time In Response. 3, 9, 10, 19, 20, 22, 23, 27, 29, 40, 43, 44

**TTR** Time To Response. 35, 36

**WHO** World Health Organization. 1

# Contents

# Chapter 1

# Introduction.

Cancer is a significant public health problem being the leading cause of death worldwide, with nearly 10 million deaths in 2020, according to the World Health Organization (WHO) [33]. Therefore, oncology clinical trials have a key role in testing new treatments, providing a scientific basis for advising and treating patients or helping scientists better understand the risk factors of the disease [19]. The WHO defines clinical trials as "*a type of research that studies new tests and treatments and evaluates their effects on human health outcomes. People volunteer to take part in clinical trials to test medical interventions including drugs, cells and other biological products, surgical procedures, radiological procedures, devices, behavioural treatments and preventive care*" [34].

In clinical trials, to allow comparison between groups, the randomization randomly allocate each patient to a group preventing a bias by distributing the characteristics of patients so if a difference appears between groups, it will only be due to the treatment [23]. The efficacy of a new treatment may then be measured through many different endpoints [30]. Some of the different endpoints most often used in oncology studies are defined by the Food and Drug Administration (FDA) [7] as follows :

- Objective Response Rate (ORR) : defined as the proportion of randomized patients whose tumor size has reduced by a predefined amount and for at least a minimum period of time.
- Response duration : defined as the time from initial response until documented tumor progression or death, whichever occurs first.
- Progression-Free Survival (PFS) : defined as the time from randomization until objective tumor progression or death, whichever occurs first. This endpoint includes the measurement of stable disease because it reflects how long the disease was stabilized and by including deaths can be a better correlate to overall survival. Different criteria for defining progression exist, it is therefore important to define tumor progression in the protocol because there are no standard regulatory criteria for progression.
- Overall Survival (OS) : defined as the time from randomization to death from any cause. This endpoint is easy to measure. Moreover, it is considered as the most reliable cancer endpoint and therefore represents the gold standard for measuring the efficacy of a new treatment by the FDA and the European Medicines Agency (EMA) [2].

The three first endpoints are based on tumor assessments performed at regular visits through imaging as MRI or PET-Scan. A classification system mostly used for solid tumors is Response Evaluation Criteria In Solid Tumors (RECIST). These criteria have been published in 2000 to determine if a cancer patient responds (tumor regresses), stabilizes (tumor size stays the same) or progress (tumor worsens) during treatment. A revision of RECIST has been introduced in 2009 as RECIST 1.1. [4] with updates to answer number of questions and issues found on the original criteria such as how to handle assessment of lymph nodes or how to use newer imaging technologies such as MRI, and so on. Thus, this guideline describes a standard approach to measure solid tumor and presents definitions for objective assessment of change in tumor size for oncology clinical trials. Globally, a response is defined as the disappearance of all lesions (Complete Response (CR)) or as a decrease of at least 30% in the sum of the lesion diameters from the sum of the initial diameters (Partial Response (PR)). On the other hand, if the sum of diameters of lesions has increased by at least 20% compared to the smallest sum observed during the study is observed and if there is an absolute increase of at least 5mm, the disease is considered to have progressed (Progression Disease (PD)). It is important to note that if one or more new lesions appear, it is also considered as a progression. Finally, if compared to the smallest sum of diameters, we cannot define PR or PD, the disease is said to be stable (Stable Disease (SD)).

Although OS is considered the gold standard endpoint in oncology clinical trials, PFS is a popular alternative endpoint because it requires fewer patients for the same study period and it allows observation of short-term progressive changes of each round of treatment. However, we must remain cautious because a prolonged PFS does not necessarily mean an extended survival [3]. On the other hand, the proportion of patients responding to a treatment and the duration of their response are also usually assessed as they represent a direct measure of anti-tumor activity of new agents. The comparison between two treatments groups (e.g. experimental versus control) for the proportion of responders is straightforward in a randomized trial but the comparison of duration of response leads to more difficulties. Indeed, analyzes that attempt to compare treatments in terms of duration of response in responder patients are likely to be biased since the groups being compared are defined by the outcome of response, i.e. a post-treatment outcome, and therefore randomization no longer holds. The subgroups of responders in each randomized treatment groups are therefore not comparable. This first problem needs to be taken into account by statisticians analyzing the data for the comparison of non-randomized groups. Moreover, a second problem exists when these two outcomes show opposite trends with a large proportion of responders but with a fairly short duration of response, or vice versa, which makes it difficult for the clinician to choose which treatment is the most effective [5]. It is then necessary to provide adequate information about this to have all information to help the clinical and patient taking informed decision with regards to treatment choice.

Three methods to analyze duration of response are studied in this master thesis : Duration Of Response (DOR) method, Time In Response (TIR) method and Probability of Being In Response (PBIR) method. This master thesis aims to review these methods, to compare them through a simulation study and to apply them in an illustration. More specifically, we will face this challenge with two objectives :

1. The main objective of this work is to compare three methods of duration of response for oncology studies, by describing the methods and the advantages and problems encountered by each, and by drawing conclusions from these methods to determine which treatment is the best.

2. The secondary objective is to determine in each method how the duration of response and the response rate will influence the decision of the best treatment and if one of these parameters is more important.

Several chapters will structure the response to these objectives. Second chapter consists in a quick overview of survival analysis concepts to introduce basic concepts used in this work. Third chapter will present the three methods studied to measure the duration of response in oncology studies. In this chapter, each method is discussed with a definition of the endpoint and its associated estimate. It is then explained how to measure the difference between treatments. Chapter 4 will present a simulation study to better understand the difference between the three methods with different situations. This chapter is sequenced in three parts: (1) the chosen simulation model is explained. This model is based on real data presented in chapter 5 ; (2) one specific (random) dataset generated in the context of this simulation study is considered and extracted for an analysis by the three methods in order to demonstrate how to interpret the results of each method ; (3) a more in-depth analysis of the three methods is carried out with a focus on the influence of the duration of response and the response rate. Chapter 5 will allow us to analyze a clinical trial with real data by the traditional endpoints and by these three methods. This chapter aims to understand in practice the different conclusions that can be drawn depending on the method used. Finally, a last chapter will include a discussion and a conclusion to close this master thesis.

Statistical analyzes are produced using *R version 4.1.1* software. The level of all statistical tests presented in this master thesis is set to 5%.

# Chapter 2

# Main concepts of survival analysis.

This chapter presenting the main concepts of survival analysis is mainly based on the following books : Kleinbaum and Klein (2011) [13] and Moore (2016) [17]. Survival analysis relates to statistical analyzes in which the outcome variable represents the time until an event occurs. This event is not limited to the death and can also be the apparition, the relapse or the progression of a disease or any experience of interest that may happen. The elapsed time is considered as the survival time, i.e. the time from the start of the follow-up to the event of interest, see figure 2.0.1 .


Figure 2.0.1: Survival analysis - Outcome variable

An important characteristic found in survival analysis is the presence of censoring. Censoring occurs when the exact time of event is not known. Here are different reasons why censoring may occur : the study ends and a patient does not experience the event before, a patient is lost to follow-up or a patient withdraws from the study for different reasons (adverse event, death or any other reason).

For example, consider the following situations (figure 2.0.2) where the event of interest is the progression of the disease or the death of the patient, whichever occurs first :

1. Patient one left the study at Month 6.
2. Patient two progressed or died at Month 5.
3. Patient three progressed or died at Month 11.
4. Patient four did not present any progression or death and quit the study at Month 9.
5. Patient five progressed or died at Month 12.

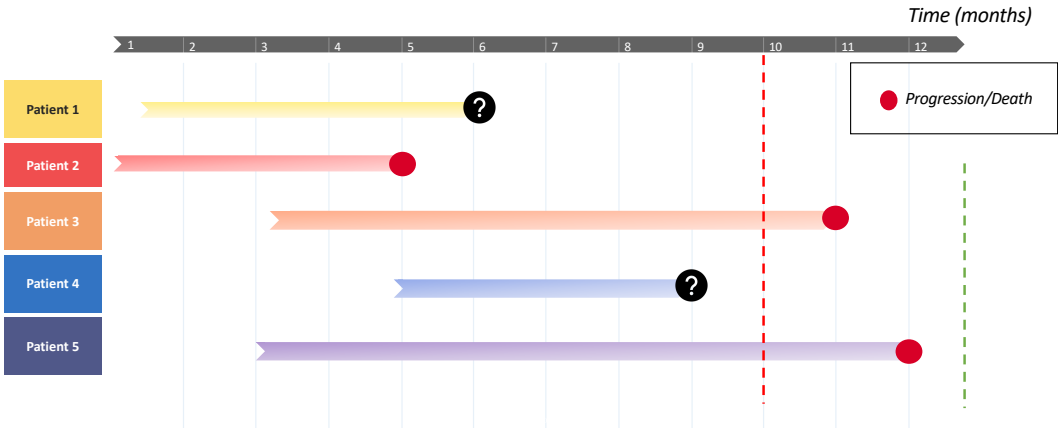


Figure 2.0.2: Censoring in survival analysis

If we consider the follow-up time equals to 13 months (green dotted line), only patient one and four are considered as censored data. Depending on the reason of the drop out, the censoring will be considered as informative or non-informative.
If the study ends at Month 10 as represented by the red dotted line, patients three and five are also considered as censored for the event of interest of progression/death. This censoring is an administrative censoring because it occurs when the study observation period ends and the patient is still in the study but does not present the event of interest. This censoring is then considered as independent, meaning that the censoring time of each patient is independent of his event time.
The presence of censoring for an individual is usually denoted by a indicator variable $\delta$ with value 1 if not censored and 0 if censored.

The type of censoring described above is called right-censoring but two other types of censoring exist : left-censoring and interval censoring. An observation is said to be left censored if the patient had the event before entering the study and interval censoring arises when the patient had the event within a time interval, which occurs if the assessment is done at a periodical frequency. In oncology studies, for time to event analysis where origin time or event time are based on tumor response as duration of response or PFS, interval censoring will occur since the tumor is only measured at fixed interval (the imaging visit). However, in practice, if the frequency of the periodicity of examination is justified and is the same for both groups, interval censoring can be treated as point censored [22]. For the rest of this master thesis, only right-censoring will be discussed as this is what is usually done in oncology trials and simply called "censoring".

Two major functions used in survival analysis are the survival function and the hazard function. Let $T$ be the time until the event of interest occurs :

- Survival function : $S(t) = P(T > t) = 1 - F(t)$ is the probability that an individual will not have the event before time $t$. $S(t)$ is decreasing and defined in [0,1].

- Hazard function : $h(t) = \lim_{\Delta t \to 0} \dfrac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$ is the conditional instantaneous

risk per unit time for the event to occur, given that the individual did not have the event at time $t$. This function is positive.

A relationship exists between these two functions as :

$$S(t) = exp\Big( - \int_0^t h(u)\, du \Big)$$

$$h(t) = -\Big( \frac{dS(t)/dt}{S(t)} \Big)$$

which allows to deduce one from the other.

A non-parametric estimation of the survival function for right censored data has been proposed by Kaplan-Meier in 1958 :

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{j:t_j \leq t} \Big[ 1 - \dfrac{O_j}{n_j} \Big] & \text{if } t \geq t_1 \end{cases}$$

where

- $0 < t_1 < t_2 < ... < t_j$ : the j distinct ordered times of event
- $O_j$ : number of events observed at time $t_j$
- $n_j$ : number of patients at risk at time $t_j$

In clinical trial, survival analysis can be used to analyze the treatment effect. Let respectively consider $S_{exp}(t)$ and $S_{ctrl}(t)$ survival functions for the experimental group and for the control group. Most of the time, the log-rank test is realized by comparing survival curves for two groups testing the following hypothesis :

$$H_0 : S_{exp}(t) = S_{ctrl}(t)\ \forall t$$

$$H_1 : S_{exp}(t) \neq S_{ctrl}(t) \text{ for some } t$$

The Log-rank test statistic follows a $\chi_1^2$ distribution under the null hypothesis. If p-value $< \alpha$, the null hypothesis is rejected and we consider that survival differs significantly between the two treatment groups.
Based on the log-rank test, it is possible to determine the sample size needed to detect a treatment effect of a certain size. The sample size depends on several parameters, including the alternative hypothesis expressed in the logarithm of the hazard ratio, the type I error and the power. In survival analysis, the power depends on the number of events and not the number of patients.

The Cox model [1] is probably the most popular regression model applied to survival data and models the effect of different variables on survival via their impact on the hazard function. A strong assumption made to use this model is the proportional hazards meaning that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time. The Cox model is written mathematically :

$$h(t|\mathbf{X}) = h_0(t) exp(\beta^t \mathbf{X})$$

where $\mathbf{X} = (X_1, X_2, ..., X_p)$ is the vector of explanatory/predictor variables and $\beta$ the vector of coefficients measuring the impact of variables. Since $h_0(t)$ is the unspecified baseline hazard function and the second part is parametric, this model is semi-parametric.

For summarizing the treatment effect, the Hazards Ratio (HR) [26] is also frequently computed estimating the ratio of event risks between two groups (under the assumption that the HR is constant over time). Let consider only one binary variable $X$ for treatment where X=1 if experimental group and X=0 if control group :

$$HR = \frac{h(t|X=1)}{h(t|X=0)} = \frac{h_0(t)exp(\beta)}{h_0(t)} = exp(\beta)$$

A HR equals to one translates an identical risk of events between the two groups. A HR larger than one (positive $\beta$) means that the risk of events is higher in the experimental group and a HR smaller than one (negative $\beta$) implies a lower risk of events for the experimental group.

# Chapter 3

# Three methods for measuring the duration of response.

This chapter presents the three methods for estimating duration of response in oncology studies: Duration Of Response, Time In Response, and Probability of Being In Response. The first two methods represent the two most commonly used methods for this outcome and the Probability of Being In Response method is a new alternative method proposed to overcome the problems encountered in these two traditional methods.

## 3.1 Duration of response.

In oncology, the Duration Of Response (DOR) is defined as the time from onset of response to disease progression or death, whichever occurs first [8] for the subgroup of patients who responded to therapy, represented in figure 3.1.1.

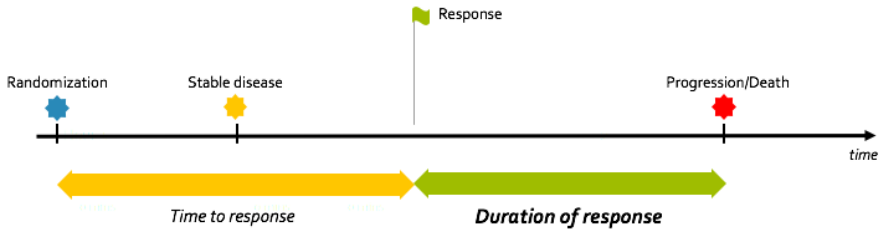


Figure 3.1.1: Duration of response

The DOR constitutes a measure of the quality, or indirectly the efficacy, of the treatments [18]. The classical approach to estimate it is using the Kaplan-Meier (KM) estimate of the survival function for DOR [16]. Some analyzes can then compare treatment arms by comparing the KM curves and the resulting median DOR between these arms. Let assume that proportional hazards of the treatment effect holds, another way to measure the difference between treatment is to compute the Hazards Ratio. Nevertheless, the DOR method presents several problems.

First of all, the consequence of ignoring non-responders can bias the assessment of duration of response, especially if the response rates of the two treatment groups differ. For example, if a treatment is considered ineffective and if few patients who are less likely to progression responded to the treatment, the mean duration of response could appear important whereas the majority of the patients did not respond to the treatment [5].

The second problem is that the factors measured after randomization (i.e. response to treatment) defining subgroups (i.e. responders vs non-responders) alter the comparison of prognostic factors (observed and unobserved) between the treatment groups, and that, despite the initial randomization. Then, using the log-rank test for comparing the distribution of time until the occurrence of an event of interest in independent groups could lead to a misleading interpretation because the randomization does not hold anymore [18].

Hence, comparison of duration of response between different treatments may be only regarded as informative and the analysis of DOR is limited to descriptive analyzes of responders (by reporting the number of responders, the number of events or the median duration of response) because only a fraction of patients responded to treatment [6] [8].

## 3.2 Time in response.

An alternative to the analysis of the duration of response is recommended by a guideline of the EMA with the Time In Response (TIR) : the non-responders would be artificially assigned a duration of response equals to zero and the duration of response for the responders is as observed [6]. All patients are therefore included in the analysis. As the DOR method, TIR is estimated by a KM estimate and its associated estimated survival function curve.

It allows us to enable a statistical comparison between groups [18] with a log-rank test and to compute a Hazards Ratio.

The major issue with that method is that the presence of artificial events for non-responders may cause difficulty to interpret the estimate and the result of the comparison between treatments.

## 3.3 Probability of being in response.

To overcome these different problems, the proposition made by Huang and al.[10] is to construct a curve from the Cumulative Response Rate (CRR), defined as the proportion of patients who have responded to treatment and remain in response at present, for the whole study period. This leads to a better visualization of the rapidity of the response to treatment, allows to see how long the response lasts and includes all patients.

The Probability of Being In Response (PBIR) represents the current response rate. The idea is then to construct a curve from the CRR which includes all the patients and considers a patient as a responder only if she/he still responds to the treatment at time $t$. The PBIR curve is then constructed by computing the difference between the two following curves in a time window:

1. Progression/Death-free Kaplan-Meier curve : representing time to progression or death, whichever occurs first. This curve is usually called the Progression-Free Survival curve.

2. Progression/Death/Response-free Kaplan-Meier curve : representing time to a composite endpoint, i.e. first of progression, death or response.

An example is presented in figure 3.3.1. On the left, P/D-free and P/D/R-free curves are displayed for one treatment group. Between these two curves, we can observe a shaded area representing the Probability of Being In Response. The PBIR curve over time can then be plotted as in the figure on the right as the difference between the two curves on the left.

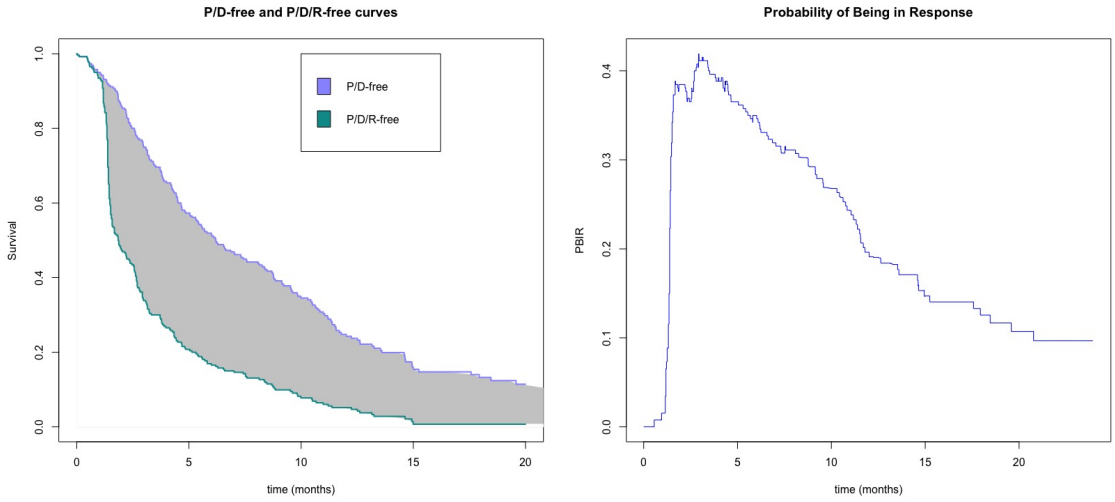


Figure 3.3.1: P/D-free and P/D/R-free curves

Values of the survival probability are extracted from the example at different times in the table 3.3.1. It can be easily observed that the difference of survival probabilities between P/D-free and P/D/R-free corresponds to the value of survival probability of PBIR.

| time (months) | **P/D-free** | **P/D/R-free** | **PBIR** |
|---|---|---|---|
| **2.96** | 0.7538 | 0.3423 | 0.4115 |
| **8.85** | 0.3914 | 0.0991 | 0.2923 |
| **10.67** | 0.3173 | 0.0646 | 0.2527 |
| **13.57** | 0.2050 | 0.0282 | 0.1768 |

Table 3.3.1: P/D-free, P/D/R-free and PBIR survival probabilities

The area under the PBIR curve represents then the mean duration of response for all patients for one treatment group. By combining the response rate and the duration of response simultaneously, using the area under the PBIR curve measures the overall treatment effect, including both responders and non-responders.

The better the treatment is for achieving and maintaining a response, the higher the PBIR curve will be. It is then possible to compare different treatments using the PBIR curves. For

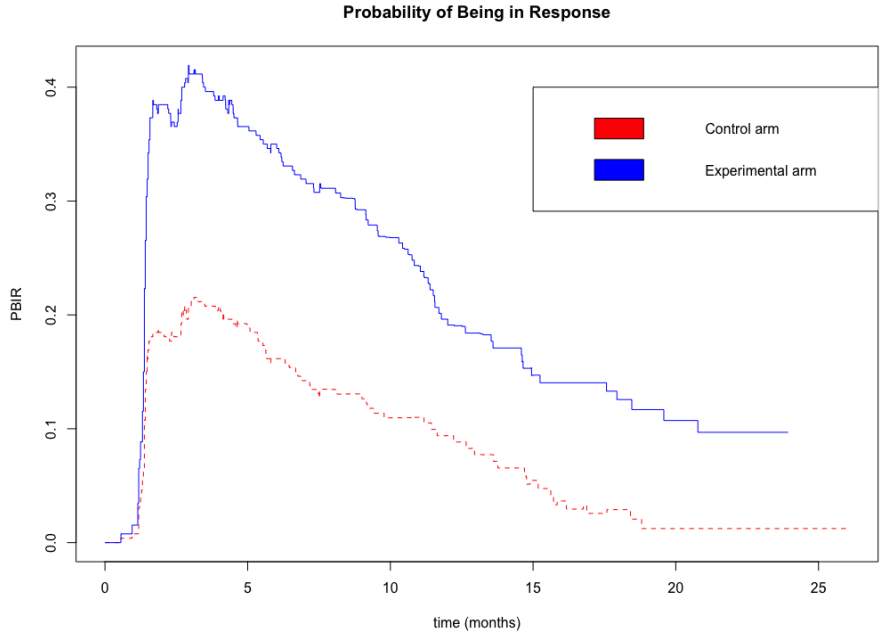


Figure 3.3.2: Probability of being in response curves comparison

example, let consider one experimental treatment compared to the standard treatment (control group) as in the figure 3.3.2. Experimental arm curve is higher than control arm curve during the whole study period meaning that proportion of responders still responding is greater for the experimental arm at each time $t$. It is also possible to compute the difference in PBIR between two groups as shown in figure 3.3.3. When the confidence interval includes the value of 0, it can be concluded that the difference in PBIR between the two treatment groups is not significant.

### 3.3.1 Method.

Two assumptions are made for computing the PBIR [29]:

- Patients can not respond after a relapse or a progression,
- Censoring time is independent of response time and progression time.

Let $X$ and $D$ be random variables, respectively time to response and time to disease progression or death (whichever occurs first). Hence, the probability of a patient of being in response (as a responder and has not progressed or died) at time point $t$ [14] :

$$PBIR(t) = P(X < t < D)$$

$$= P(D > t) - P(X > t)$$

Let $Y$ be a random variable of the minimum between time to response and time to disease progression or death (i.e. $Y = min(X, D)$). Then,

$$PBIR(t) = P(D > t) - P(Y > t)$$

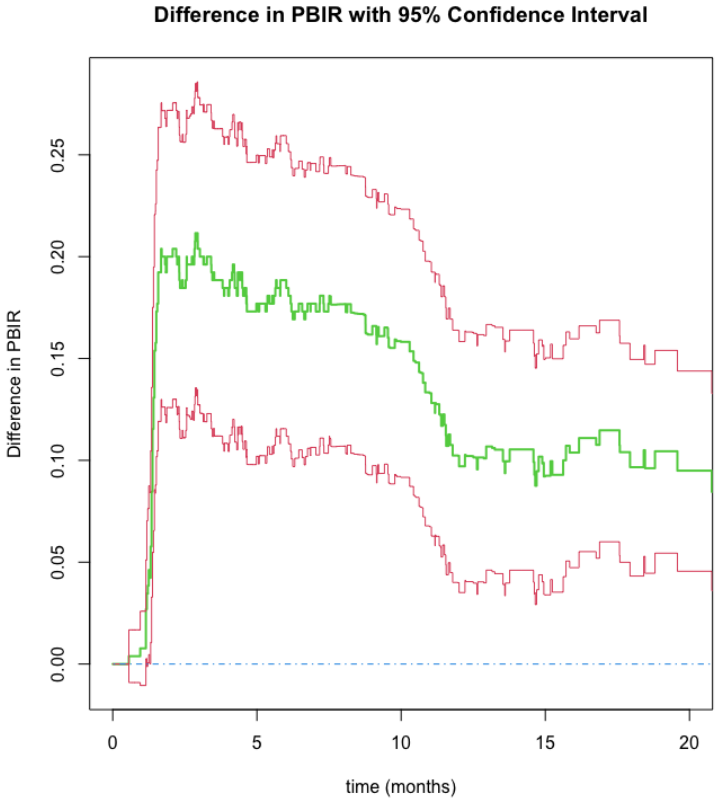


Figure 3.3.3: Difference in PBIR between two groups

Non-parametric estimator of the PBIR is the difference between two Kaplan-Meier estimates for the survival function of $D$ and $Y$ in the presence of censoring :

$$\hat{PBIR}(t) = \hat{S}_D(t) - \hat{S}_Y(t)$$

To summarize the information provided by the PBIR curve into a quantity, it is possible to calculate the mean duration of response (which can also be used for inference) defined as :

$$E[(D - X)I(D > X)] = E[D - Y]$$

representing the expected duration of a patient being a responder where $E[.]$ is the expected value and $I(.)$ is an indicator function of an event taking value 1 when the event happens and value 0 when the event does not happen. For a patient who does not respond to treatment, the mean duration of response is zero. Within a time window $[0, \tau]$ where $\tau$ is a time point no more than the longest follow-up time in the data, the mean duration of response is estimated as :

$$E[min(D, \tau) - min(Y, \tau)] = E[min(D, \tau)] - E[min(Y, \tau)]$$

Where

$$E[min(D, \tau)] = \int_0^{+\infty} min(u, \tau) f_D(u) du$$

with $f_D$ the probability density function of $D$.

But, $min(D, \tau) = \int_0^\tau I[D > t]dt$.

Then,

$$\begin{aligned} E[min(D, \tau)] = \int_0^{+\infty} min(u, \tau) f_D(u) du &= \int_0^{+\infty} \int_0^\tau I[d > t] dt f_D(u) du \\ &= \int_0^{+\infty} \int_0^\tau I[d > t] f_D(u) dt du \\ &= \int_0^\tau \int_0^{+\infty} I[d > t] f_D(u) du dt \\ &= \int_0^\tau P(D > t) dt \\ &= \int_0^\tau S_D(t) dt \end{aligned}$$

In the same way,

$$E[min(Y, \tau)] = \int_0^\tau S_Y(t) dt$$

Thus,

$$E[min(D, \tau) - min(Y, \tau)] = \int_0^\tau S_D(t) dt - \int_0^\tau S_Y(t) dt$$

Alternatively, this expression can be expressed as the area under the PBIR curve :

$$\int_0^\tau PBIR(t)\, dt$$

We can then observe that the mean duration of response can be estimated as the difference of two integrals of Kaplan-Meier curves :

$$\int_0^\tau P\hat{B}IR(t)\,dt = \int_0^\tau \hat{S}_D(t)dt - \int_0^\tau \hat{S}_Y(t)dt$$

The integral of one KM curve can then be estimated by the Area Under Curve (AUC) by summing the areas of the sub-rectangles under the KM curve created at each event time. For example, using the figure 3.3.4, the AUC of this survival function is equal to :
$AUC = (1*10) + (0.85*(25-10)) + (0.6*(35-25)) + (0.55*(44-35)) + (0.42*(60-44)) + (0.24*(73-60)) + (0.21*(82-73)) + (0.1*(100-82)) = 47.23$ days.

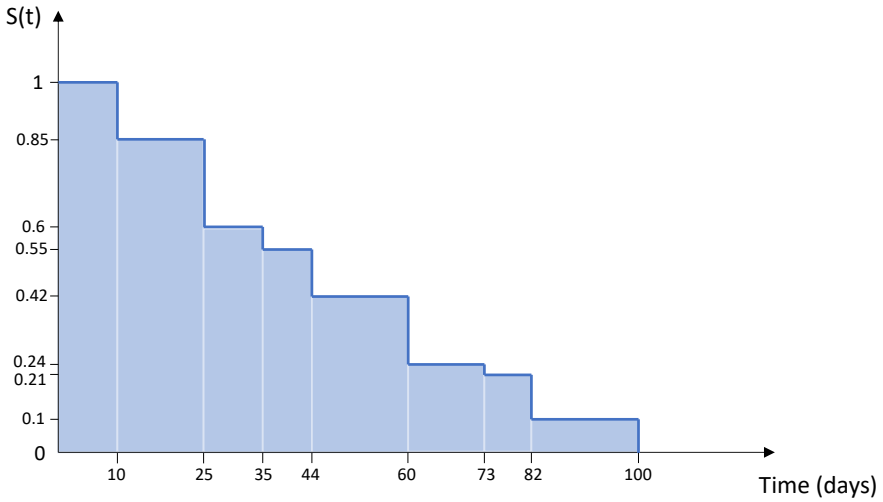


Figure 3.3.4: AUC computation

The difference in mean duration of response between groups is computed by subtracting the mean duration of response of one group from the one of the other group. The standard error of this difference can also be computed as $SE = \sqrt{SE_1^2 + SE_2^2}$ where $SE_1^2$ and $SE_2^2$ are respectively the standard error of mean duration of response of group 1 and group 2. It is then possible to constructed a curve of difference in mean duration of response between groups with confidence intervals.

Additionally, the Wald Test [32] can be used as a statistical test to compare mean duration of response between two groups in the same time window with the test statistics and its corresponding p-value. Also called the Wald Chi-Squared Test, this test assesses whether explanatory variables in a model are significant with the following hypothesis :

$$H_0 : \beta = \beta_0$$

$$H_1 : \beta \neq \beta_0$$

where $\beta$ is the coefficient of the explanatory variable and $\beta_0$ is the parameter of interest, usually equals to 0 to test if the coefficient is different than zero. Test statistic equals to :

$$\chi^2_W = (\hat{\beta} - \beta_0)^t \mathbf{I}(\hat{\beta})(\hat{\beta} - \beta_0) \sim_{H_0} \chi^2_1$$

where $\hat{\beta}$ is the maximum log-likelihood estimator and $\mathbf{I}(\hat{\beta})$ the observed information matrix. The p-value can be computed as $P(\chi_1^2 \geq \chi_W^2)$.

# Chapter 4

# Simulation study.

A simulation study is performed to apply the three methods studied to fictitious data in order to observe the behavior of the conclusions drawn by each method. In this chapter, a first section details the chosen simulation model and the implementation in the software *R*. The second section includes the analysis of a dataset extracted from the simulations to directly compare the three methods for duration of response and to demonstrate how to interpret the results in practice. The last section contains a more in-depth analysis of the three methods to observe the influence of two parameters, the response rate and the duration of response, on the results of each method.

## 4.1 Simulation Model.

The goal of this simulation study is to compare the conclusions of the three methods for measuring the duration of response in settings with different response rates and different durations of response for the experimental arm compared to control arm. The simulation model used is presented in figure 4.1.1.

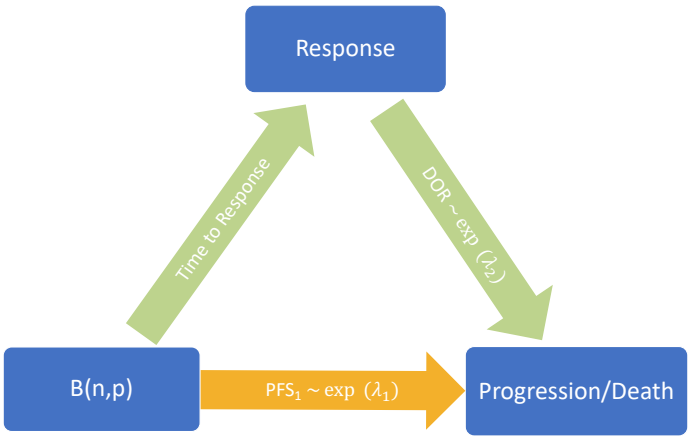


Figure 4.1.1: Simulation model

| Parameters | Value |
|---|---|
| $p_{ctrl}$ | 0.3077 |
| $\lambda_{1ctrl}$ | 0.0067 |
| $\lambda_{2ctrl}$ | 0.0038 |

Table 4.1.1: Values of control arm parameters of the simulation study

For beginning, a binomial distribution with parameters $n$ (representing number of patients in one group) and $p$ (representing response rate) is used to determine each patient as a responder or a non-responder. If the patient is a responder, he will get a time to response and a duration of response, the latter following an exponential distribution with parameter $\lambda_2$. His total time to progression or death will be the sum of his time to response and his duration of response. If the patient does not respond, he will get a time to progression or death, whichever occurs first, following an exponential distribution with parameter $\lambda_1$. As a reminder, if $T$ (time to event) follows an exponential distribution with parameter $\lambda$, the survival function is : $S(t) = exp(-\lambda t)$.

To select the parameter values, we decide to follow the real data case presented in more details in chapter 5 and to mimic the control arm data. Therefore, 520 patients are enrolled in each trial ($n = 260$) and time to response is simulated by resampling time to response of responders of the control arm (figure 5.0.2 in Chapter 5). To simulate the control arm of the simulation study, the simulation model is applied and we rely on the Maximum Likelihood Estimator on this data for setting $p_{ctrl}$ from a binomial distribution and $\lambda_{1ctrl}$ and $\lambda_{2ctrl}$ from two exponential distributions. Values of these parameters can be found in table 4.1.1. Note that value of $\lambda_1$ is fixed to the same in both arms, assuming that there is no impact of the treatment on the PFS of non-responders. Based on these values, simulations settings are defined by modifying the response rate and the duration of response of the experimental arm compared to control arm as :

- Modifying the response rate for experimental arm with $p_{exp} = p_{ctrl} + p_{diff}$ where $p_{diff}$ can take the following values from detrimental effect in response rate to positive effect : $-0.25, -0.2, -0.15, -0.1, -0.05, 0, 0.05, 0.1, 0.15, 0.2, 0.25$.
- Modifying the duration of response for experimental arm with $\lambda_{2exp} = \lambda_{2ctrl} * \lambda_{2diff}$ where $\lambda_{2diff}$ can take the following values from detrimental effect in duration of response to positive effect : $1.5, 1.4, 1.3, 1.2, 1.1, 1, 0.9, 0.8, 0.7, 0.6, 0.5$.

If $p_{diff}$ takes value 0 and $\lambda_{2diff}$ takes value 1, it corresponds to a setting where experimental and control arms have the same response rate and the same duration of response (i.e. no treatment effect).

Finally, the censoring status is created. Figure 4.1.2 schematizes the different situations of censoring. For each patient, a time of his entry into the study is assigned by a uniform distribution from 0 to $A$ (representing the accrual time) and is added to the total time until progression or death, whichever occurs first. If this time is greater than the sum of the study accrual time and the study follow-up ($mFU$) time , the patient is then considered to be censored (as patients 3 and 4) at the end of the study follow-up time. $A$ and $mFU$, defined based on the phase III study of chapter 5, take respectively value of 20 months and 7 months.

Thus, the created censoring is called an administrative censoring because this right-censoring occurs when the event is not observed while the observation study period is over (represented by the dotted red line) [27] and does not depend of response time.

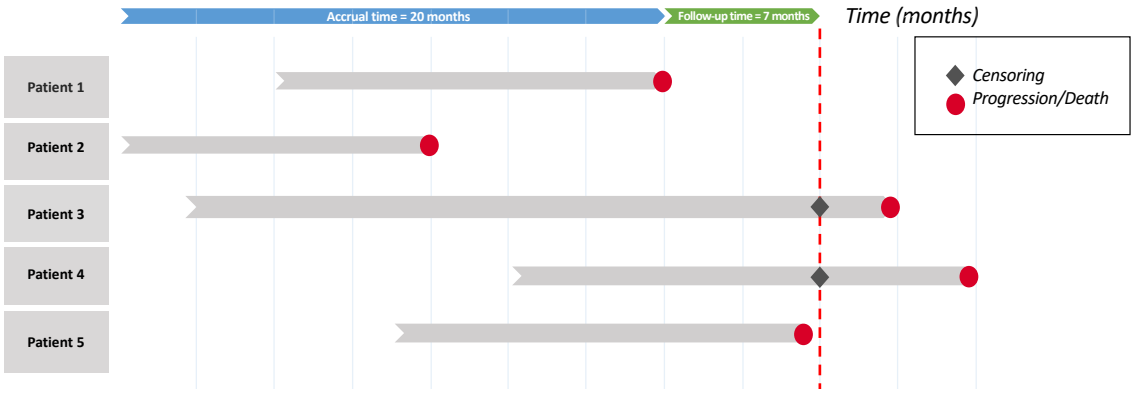


Figure 4.1.2: Censoring of the simulation model

For each pair of $p_{diff}$ and $\lambda_{2diff}$, 1000 replicates of trial are performed and comparison of the two arms is made by : (1) DOR method with hazard ratio between groups with associated p-value of log-rank test; (2) TIR method with hazard ratio between groups with associated p-value of log-rank test ; (3) PBIR method with difference in mean duration of response with associated p-value of Wald test. For summarizing the result of each setting, the means of the 1000 replicates for HR of DOR, for HR of TIR and for difference in mean duration of response (PBIR) with the maximum time $\tau$ for computing mean duration of response fixed to the percentile 95 for avoiding variations at last observed time, and proportions (for each method) of trials with a significant p-value (i.e. p-value $\leq 0.05$) of the 1000 replicates are computed. Moreover, three other outputs are measured to complete the analysis :

- The mean over the 1000 censoring rates of the whole dataset (control and experimental arms).
- The mean over the 1000 estimated variances of difference in mean duration of response between groups (by the PBIR method).
- The mean over the 1000 Odds Ratios between groups are computed and a Fisher test with the corresponding p-value is used to compare the obtained OR to value one. A odd represents the probability of success over the probability of failure, (e.g. the probability of response over the probability of non-response). The OR can be estimated as :

$$\widehat{OR} = \frac{\widehat{odd}_{exp}}{\widehat{odd}_{ctrl}} = \frac{\frac{\hat{\pi}_{exp}}{1-\hat{\pi}_{exp}}}{\frac{\hat{\pi}_{ctrl}}{1-\hat{\pi}_{ctrl}}}$$

  where $\pi$ is the probability of success (e.g. response). The Fisher test evaluates if there is an independence between the group (experimental or control) and the response (yes or no) as the null hypothesis. An alternative to this test is to perform a Chi-square test, but with a large enough sample, the results of both tests are similar. An OR greater than one correspond to a higher odd of being in response in experimental group.

### 4.1.1 Implementation in R.

Firstly, the DOR method is implemented through the **survival** package with the *coxph* function to obtain the p-value of the log-rank test as well as the estimate of the $\beta$ coefficient. The exponential of this coefficient estimates the HR.

Secondly, after the creation of the time in response variable and a new censoring variable, the package **survival** is used with the function *coxph* to obtain the p-value of the log-rank test and the HR for the TIR method.

Thirdly, the output from the PBIR method is obtained by using the **PBIR** package from Luo and al. [14]. This package contains four different functions to calculate the PBIR curve over a given time window, to compare PBIR over a time window between two groups, to do inferences on the mean duration of response and to do inferences on CRR. In this simulation study, we used the *mduration* function to compute difference in PBIR between groups, the variance and the p-value of the Wald test. Plots of the estimated PBIR curves of two treatment groups over a time window $[0 ; \tau]$ are produced by the *PBIR1* function and the difference in PBIR between groups with confidence intervals by the *PBIR2* function.

Finally, the *fisher.test* function from the **stats** package allows us to obtain the p-value of the fisher test and the estimate of the OR by the logarithm of the obtained estimate.

The complete code for this work can be found on the following link :

https://github.com/thuyldo/master_thesis.git

## 4.2 Example of data analysis for response outcomes with $\lambda_{2diff} = 0.8$ and $p_{diff} = 0.1$.

One of the datasets created by the simulation study is extracted in order to observe the different response measures and to illustrate the different conclusions of each method. The analyzed dataset presents as fixed parameters $\lambda_{2diff} = 0.8$, $p_{diff} = 0.1$ and $\lambda_{1diff} = 1$, which represents a higher response rate of 10% and a longer duration of response for the experimental group. With the chosen setting, we expected to observe a significant difference between groups and this section aims to observe if this difference is detected by each method. Time to response resulting from this simulation is shown in the figure 4.2.1.

First of all, the results of the DOR method are shown in the figure 4.2.2-A. The median duration of response for experimental group is 10 months (IQR [3.74 - NA]) vs 7.91 months (IQR [3.09 - 16.45]) for control. By a Cox PH regression, the resulting HR is 0.8233 (95% CI [0.5773 ; 1.174]), close to the value of $\lambda_{2diff}$ of 0.8 as expected, and the log-rank test shows a p-value=0.2823, indicating a non-significant difference between the two treatment groups.

Then, the TIR demonstrates a HR of 0.7695 (95% CI [0.6415 ; 0.923]) with a p-value corresponding to the log-rank test of 0.0046, showing a significant difference between the two groups. We can observe in the figure 4.2.2-B a higher drop for the control group than for the experimental group on the first day, obtained by adding artificial events for the non-responders.

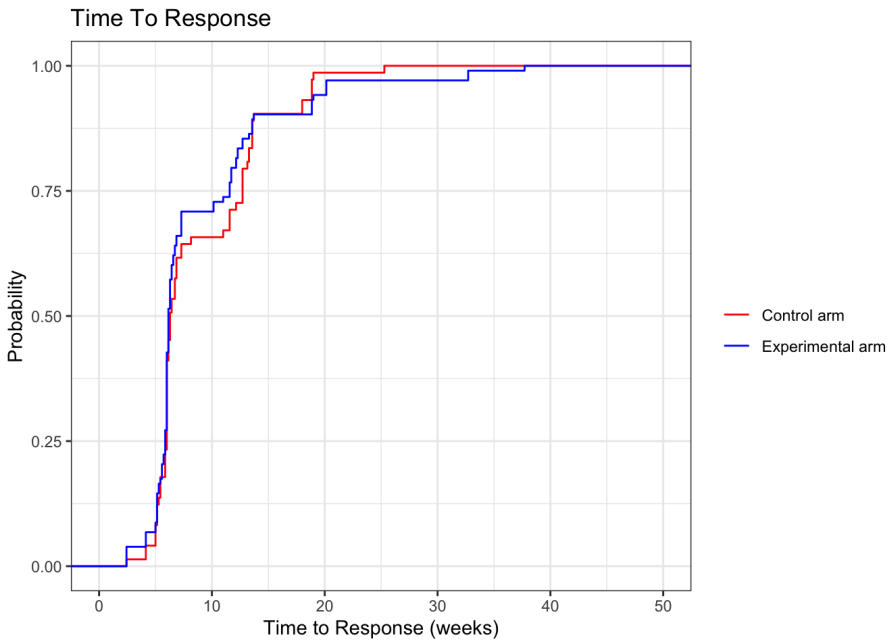


Figure 4.2.1: Simulation dataset - Time to Response

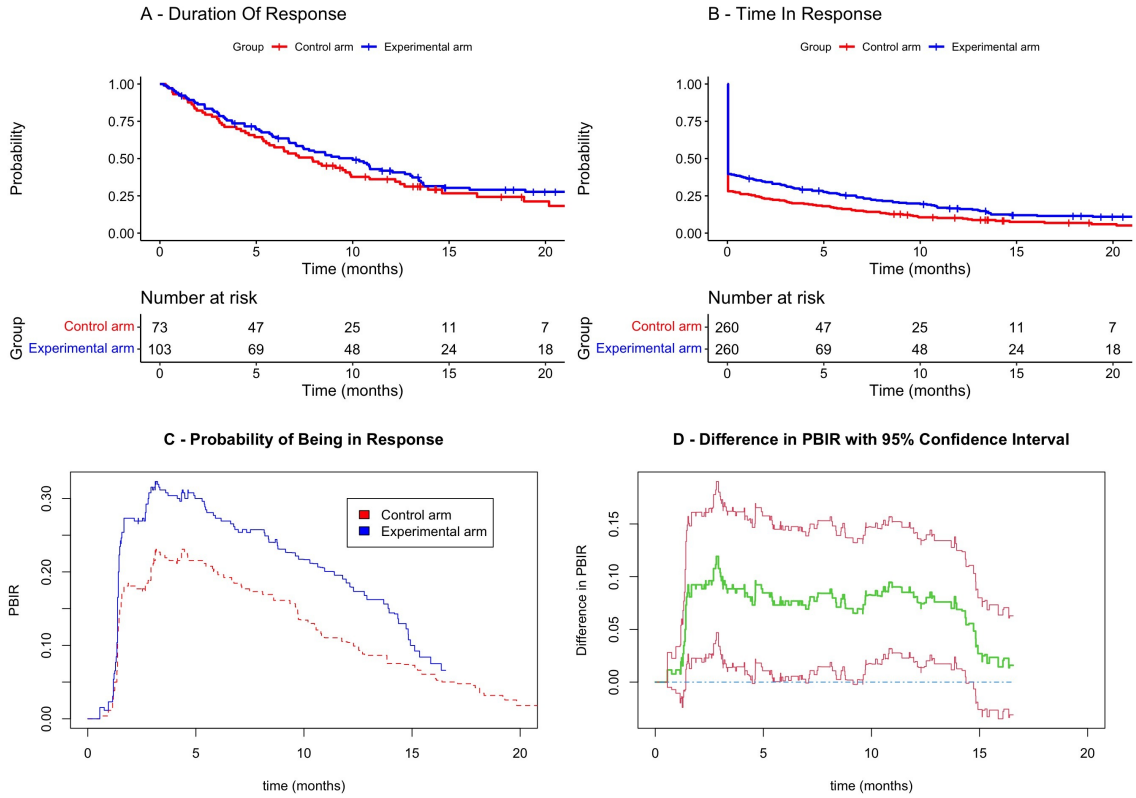


Figure 4.2.2: Simulation dataset ($\lambda_{2diff} = 0.8$ and $p_{diff} = 0.1$) - Resulting plots for three methods

Proportions of responders are respectively 33.08% and 47.69% for control and experimental groups. The median time in response is similar between groups but the 3rd quartile amounts to 6.67 months for the experimental group whereas it is only 1.44 months for the control group.

Finally, as the blue curve is higher, the PBIR shows a larger difference between the two KM curves (P-D-free and P-D-R free) for the experimental group (figure 4.2.2-C). This larger difference is confirmed in the figure 4.2.2-D, where the estimate of the PBIR difference lies above 0. The difference between two groups seems larger from month 2 to month 10. Moreover, the inference on the mean duration of response shows a p-value of 0.0309, signifying a significant difference between groups at the 5% level. This significant difference of mean duration of response between groups is equal to 0.9632 months (95% CI [0.0884 ; 1.8381]) over the time window [0 ; 16.6 months].

A summarized table with results are presented in table 4.2.1. By comparing these three methods on this dataset, the DOR method considering only the responders did not find a significant difference between the groups, while the other two methods show a significant difference between the experimental and control groups, difference that we forced to exist by fixing the parameters of $\lambda_{2diff}$ and $p_{diff}$. By comparing the TIR and the PBIR methods, we realize that the TIR gives a very low p-value (i.e. equal to 0.005), with an HR of 0.77, below the imposed HR of 0.8 in this example. The PBIR meanwhile demonstrates a significant difference for a level of 5% with a p-value of 0.0309.

| | *Estimate* | *Confidence interval* | *Associated p-value* |
|---|---|---|---|
| **DOR** | HR = 0.8233 | [0.5773 ; 1.174] | 0.2823 |
| **TIR** | HR = 0.7695 | [0.6415 ; 0.923] | 0.0046 |
| **PBIR** (months) | MDOR = 0.9632 | [0.0884 ; 1.8381] | 0.0309 |

Table 4.2.1: Simulation dataset - Results for three methods

## 4.3 Results.

We are now interested in analyzing how variations of duration of response and response rate impact the results of these three methods, based on our simulations. For this, $\lambda_{2diff}$ and $p_{diff}$ have taken 11 different values, respectively between 0.5 and 1.5 and -0.25 and 0.25.

Firstly, Duration Of Response is assessed with the Hazards Ratio and the p-value of log-rank test. These results are shown in figure 4.3.1. As observed, the HR does not vary with $p_{diff}$, thus the only parameter that influences this response outcome is the duration of the response. When there is no difference between two groups (i.e. $\lambda_{2diff} = 1$ and $p_{diff} = 0$), the HR equals to one and this value becomes greater when $\lambda_{2diff}$ increases, corresponding to a duration of response shorter in the experimental group and meaning a higher risk of progression or death in this group. Numerical results are presented in appendix A.1. Note that, as expected, HR is very close to $\lambda_{2diff}$. Furthermore, in the proportion of trials with a significant p-value, the difference is the result of variations of $\lambda_{2diff}$. The further away we get from $\lambda_{2diff} = 1$, the greater is the proportion of trials showing a significant difference between treatments. The power of a test, defined as the probability of detecting a difference between treatments when the null hypothesis is false, depends on the log(HR) [25]. The variance of the

estimated log(HR) is equal to $4/d$ where d represents the number of events [21]. Therefore, the power depends on $d$, the number of events. The asymmetrical shape observed and the difference between the curves of $p_{diff}$ are thus due to the number of events observed. Indeed, if $p_{diff}$ is high, meaning a higher number of responders, we will expect a greater value of $d$ for a same $\lambda_{2diff}$ and same follow-up, hence the power will be higher.

Secondly, Time In Response method demonstrates variations of HR with different values of duration of response and different values of response rate (figure 4.3.2 and appendix A.2). On one hand, the increase of response rate for experimental group gives a smaller HR for a same value of $\lambda_{2diff}$. In the same way, a longer duration of response for experimental group ($\lambda_{2diff}$ smaller) gives a smaller HR for same value of $p_{diff}$. Thus, in this method, Hazards Ratio is influenced by the two parameters. For each value of $p_{diff}$, the variation of $\lambda_{2diff}$ seems to influence the HR in the same way giving parallel curves. But for the same value of $\lambda_{2diff}$, a smaller $p_{diff}$ influences more strongly the HR, shown by a different gap between the curves. Indeed, let take $\lambda_{2diff} = 0.6$ for example. Between values of $p_{diff}$ of 0.25 and 0.2, the difference in HR is only 0.06 (0.60 - 0.54) when the difference amounts to 0.28 between values of $p_{diff}$ of -0.2 and -0.25. On the other hand, we can observe that the proportion of trials with a significant difference between groups varies according to both the duration of response and the response rate. When no difference for $p$ and $\lambda_2$ exists between the groups, this proportion amounts to 5%. When $p_{diff} \geq 0.2$ and when $p_{diff} \leq -0.15$, the proportion of trials with a significant difference is always greater than 50%, regardless of the value of $\lambda_{2diff}$. Although, independently of $p_{diff}$, no value of $\lambda_{2diff}$ allows to have a majority of trials with a p-value higher than 0.05.

Thirdly, results for Probability of Being In Response method are presented in figure 4.3.3 and numerical results can be found in appendix A.3. As expected, no difference in PBIR (i.e. 0 days) is observed with $p_{diff} = 0$ and $\lambda_{2diff} = 1$. Furthermore, from the figure we can see that both parameters affect the difference in PBIR between the two groups. When the number of responders increases in the experimental group, the difference between groups is larger. In the same way, for the same value of $p_{diff}$, when the duration of response increases for the experimental arm, the difference in PBIR also increases, and this is even more important when the response rate is large. For example, we can observe that if $p_{diff} = 0.1$, the range of difference in PBIR according to the value $\lambda_{2diff}$ is only 46 days (48.32 - 2.32) whereas when $p_{diff} = 0.25$, this range is almost 67 days. The results for proportion of trials with significant p-value show a plot with a similar form to the plot of the TIR method seen before. Positive values of $p_{diff}$ give decreasing curves while negative values show increasing curves, meaning that for a positive $p_{diff}$, a difference between groups is more marked when $\lambda_{2diff}$ is small and for negative $p_{diff}$, this difference is more observed with a large $\lambda_{2diff}$.
To further understand the influence of duration of response and response rate on the difference in PBIR, we can analyze the contour plot in figure 4.3.4. We observe that a value of $p_{diff} = 0.1$ is sufficient to obtain a positive difference in PBIR for all values of $\lambda_{2diff}$ (between 0.5 and 1.5). The lower is the value of $p_{diff}$, the lower is the slope of the curve and the lower is the influence of the duration of response on the PBIR difference. In addition, we see that no value of $\lambda_{2diff}$ can guarantee a positive difference in PBIR between the two groups.

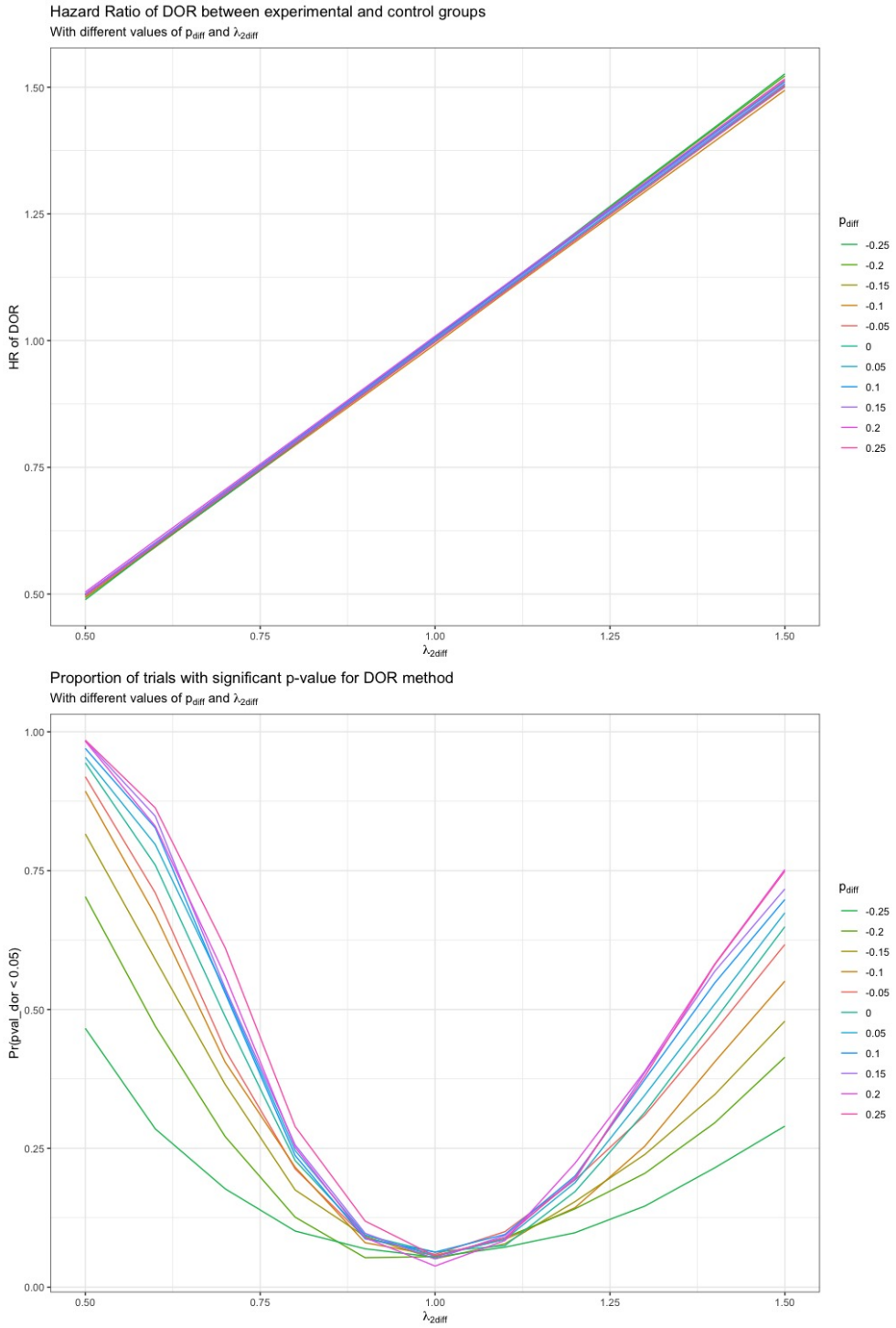


Figure 4.3.1: Simulation results for Duration of Response

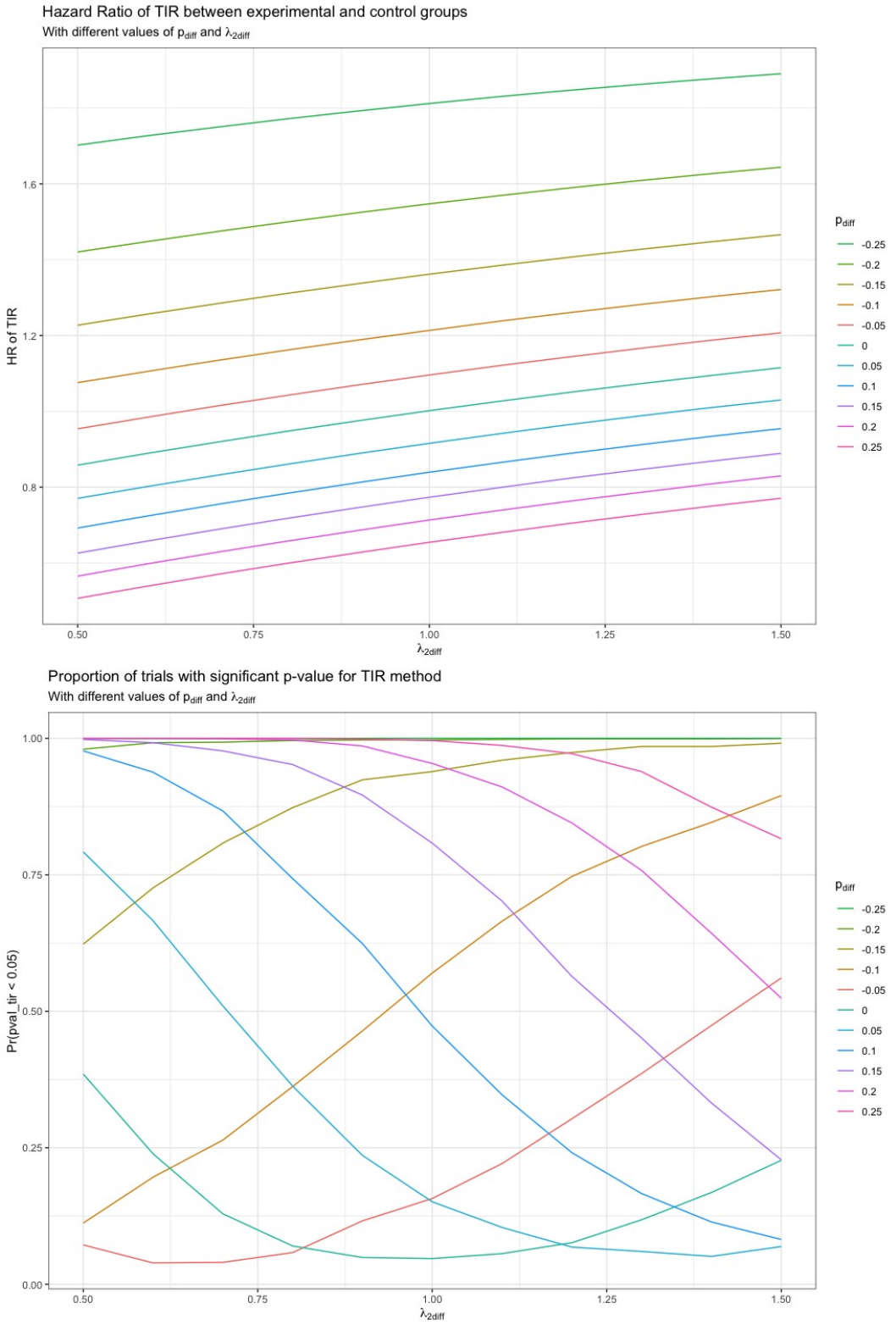


Figure 4.3.2: Simulation results for Time In Response

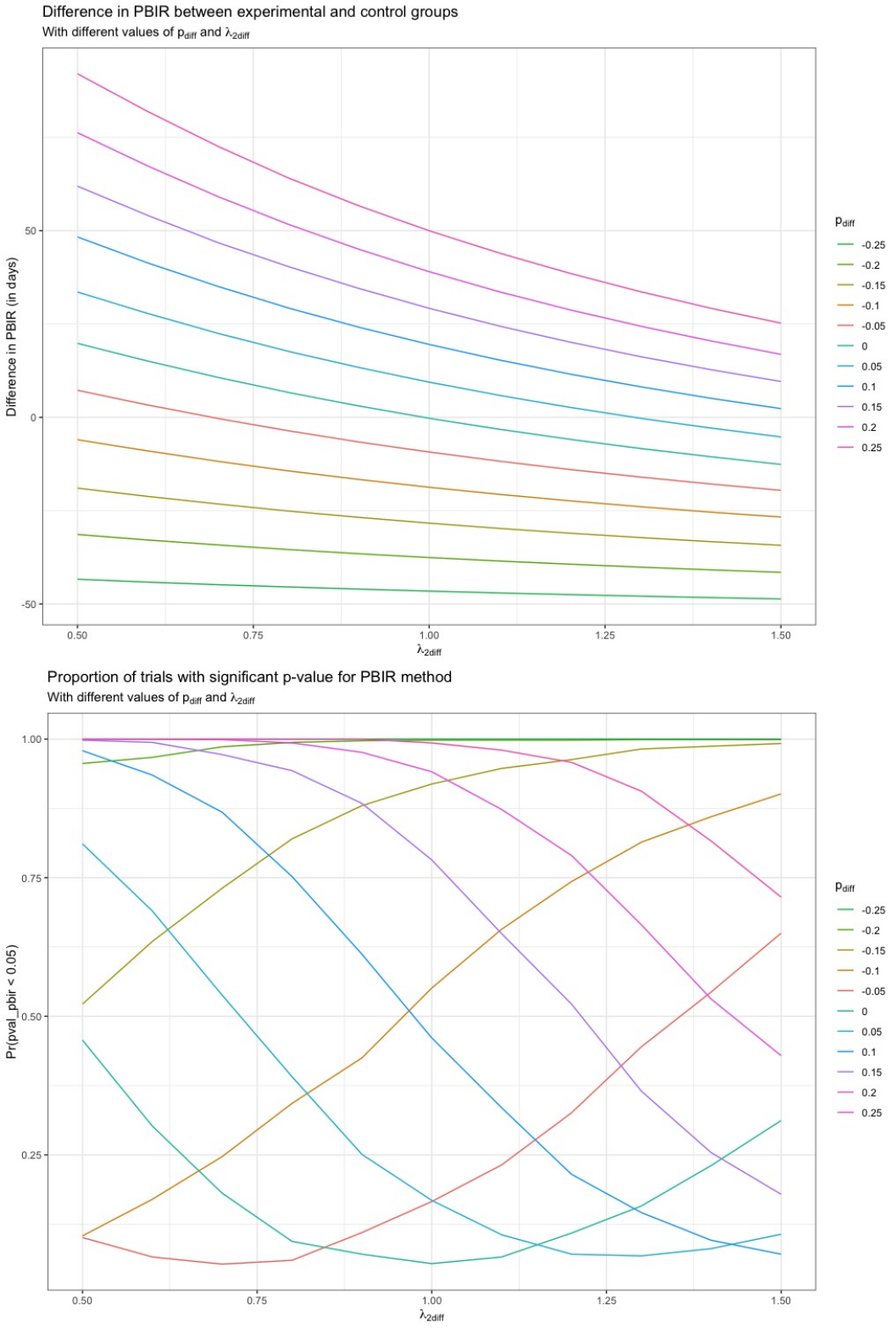


Figure 4.3.3: Simulation results for Probability of Being In Response

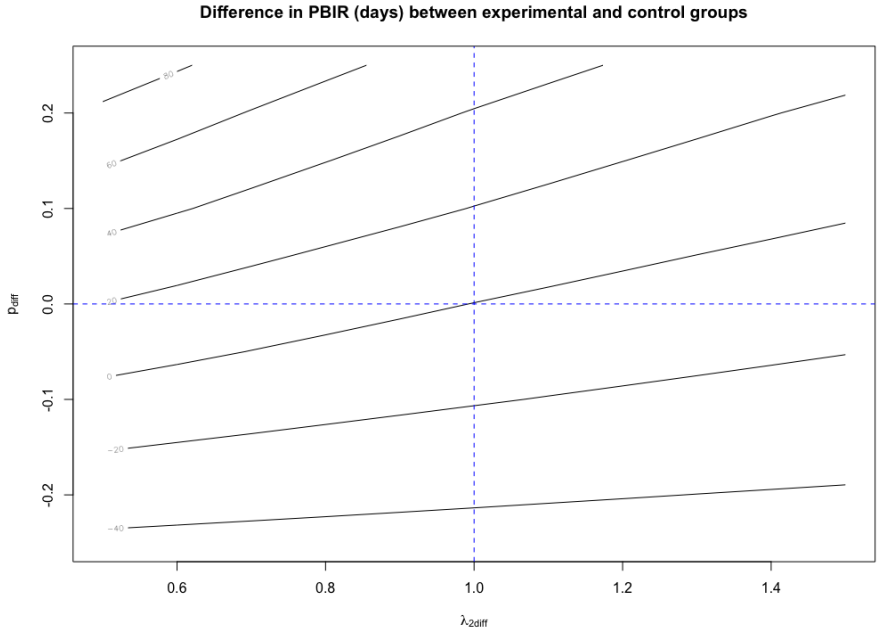


Figure 4.3.4: Simulation results for Probability of Being In Response - Contour plot

Moreover, to compare the different methods on the proportions of trials with a significant difference, the results with $p_{diff} = 0.15$ are extracted from each method and gathered in the figure 4.3.5. Let take the example to have a majority of trials showing a p-value lower than 0.05, the DOR method shows the necessity to have a $\lambda_{2diff}$ lower than 0.75 (treatment arm is better) or higher than 1.3 (control arm is better) due to its parabolic shape. The TIR and PBIR methods show a decreasing curve where it is necessary to have a value lower than 1.25 for the TIR and a value little smaller for the PBIR to have more than 50% of the trials with a significant difference where the treatment effect still positive but decrease with larger value of $\lambda_{2diff}$. These two methods therefore cover a wider range of $\lambda_{2diff}$ values than the DOR method to show a defined proportion of trials with a significant difference between groups. Note that the test conducted is two-sided, so very large values of $\lambda_{2diff}$ are required to observe the increasing part of the curve (where the significant difference between treatments is due to a better control arm) for TIR and PBIR methods.

Mean censoring rates of the 1000 simulated datasets are presented in figure 4.3.6. For the central value, the censoring rate stands at 10.74% (appendix A.4). This rate varies with the values of $p_{diff}$ and $\lambda_{2diff}$ because the censoring status was created as an administrative censoring. As a reminder, when the time to progression exceeds the time to $A + mFU$, then the patient is censored. A longer duration of the response (i.e. small $\lambda_{2diff}$) or a higher response rate (i.e. high $p_{diff}$) increases thus the number of censored observations. Indeed, the smallest censoring rate (8.45%) happens when $p_{diff} = -0.25$ and $\lambda_{2diff} = 1.5$ while the highest censoring rate (19%) happens with $p_{diff} = 0.25$ and $\lambda_{2diff} = 0.5$. The results for the variance of the PBIR difference shows a similar graph (figure 4.3.7) as the censoring rate and appears to be impacted by the number of events and indirectly by the censoring rate. The larger is the censoring rate, the higher is the variance.

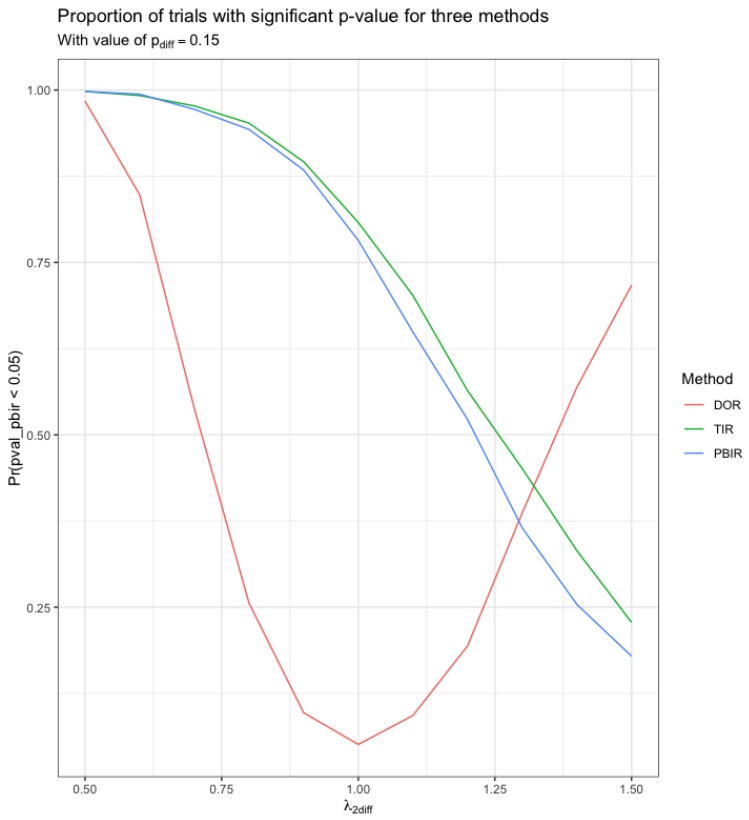


Figure 4.3.5: Simulation results for proportion of trials with significant p-value - Methods comparison

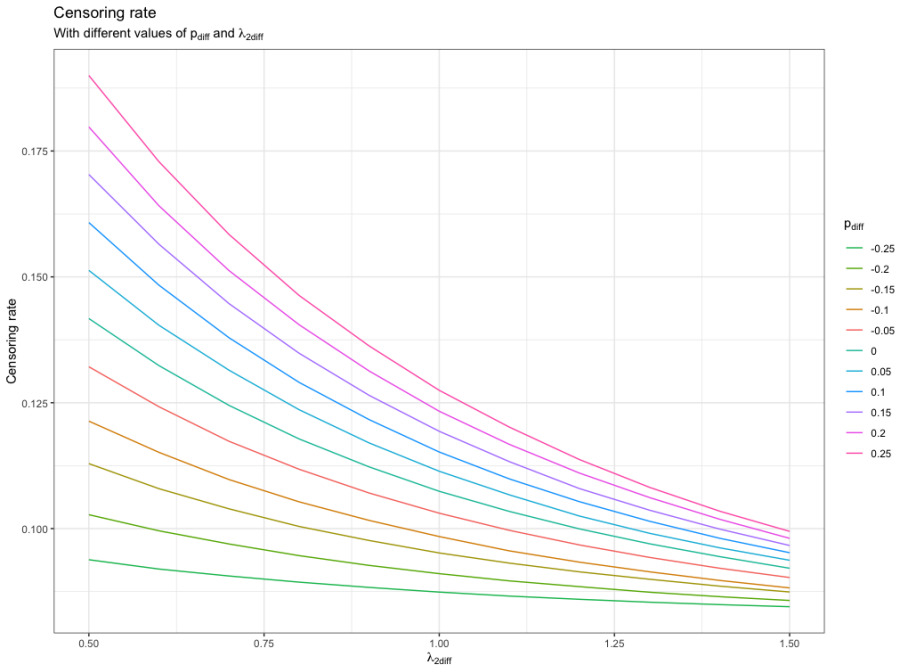


Figure 4.3.6: Simulation results for censoring rate

Furthermore, results obtained with the Odds Ratio and the Fisher Test seem consistent (figure 4.3.8 and appendix A.5). When there is no difference in response rate and response duration, the OR obtained is equal to 1. The OR tends to decrease when the response rate of the experimental group is smaller than the one of the control group meaning a lower probability of response in the experimental group and vice versa when $p_{diff}$ is greater than 0. Moreover, as expected, the OR does not depend on the value of the duration of the response. Concerning the obtained p-values, the proportion of trials with p-value under 0.05 is also not related to the value of $\lambda_{2diff}$ and increases when $p_{diff}$ deviates from 0, as shown in the figure.

Finally, for completing this analysis, we were interested if the value of $\lambda_{1diff}$, i.e. the time to progression or death for non-responders, could have an influence on these methods. So, simulations are made by modifying the time to progression or death for experimental arm with $\lambda_{1exp} = \lambda_{1ctrl} * \lambda_{1diff}$ where $\lambda_{1diff}$ can take the following values from detrimental effect to positive effect :

$$1.5, 1.4, 1.3, 1.2, 1.1, 1, 0.9, 0.8, 0.7, 0.6, 0.5$$

Parameters $p_{diff}$ and $\lambda_{2diff}$ are respectively fixed arbitrarily at 0.15 and 0.8, meaning a higher response rate and longer duration of response in the experimental arm. Obviously, the HR from the two first methods (DOR and TIR) do not change with variations of $\lambda_{1diff}$ (figure 4.3.9). The value of HR stands at 0.801 for DOR method and at 0.718 for TIR method.

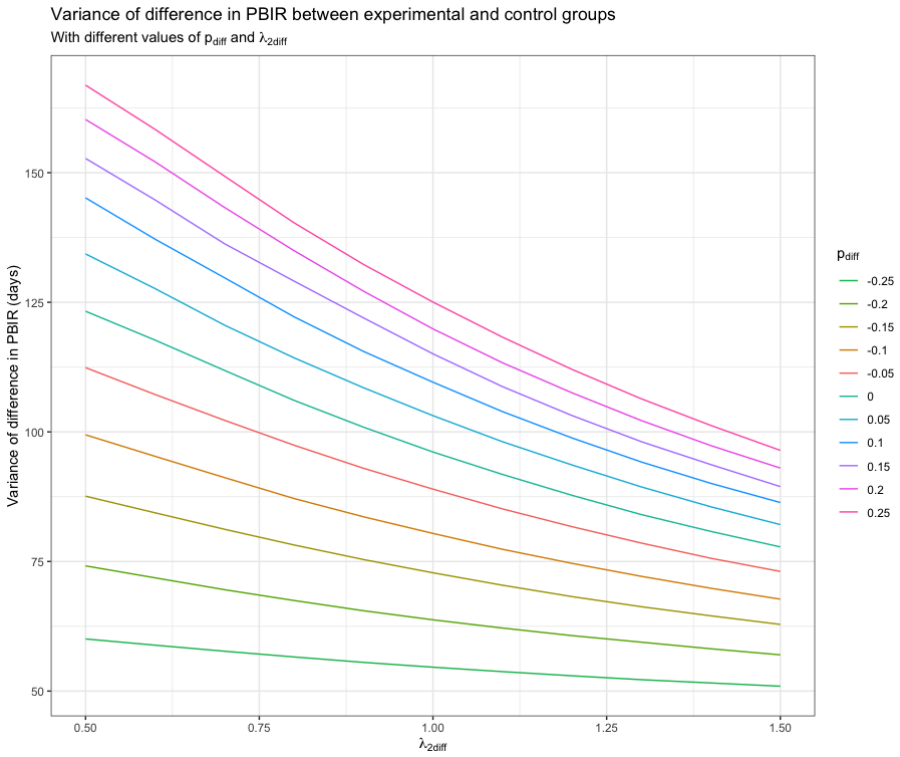


Figure 4.3.7: Simulation results for variance of difference in PBIR

For PBIR method, $\tau$ was fixed to the percentile 95 of the longest follow-up time for responders with an event to avoid missing too much information of the experimental arm. We expect the same result than other methods but based on the simulation study, difference in PBIR between groups appears different according to the value of $\lambda_{1diff}$. To better understand, we decide to vary the follow-up time $mFU$ (and thus to modify the censoring rate) which can take the following values (in months) : $1, 3, 5, 7, 11, 13, 100$. Therefore, when the follow-up time is large enough to reduce the censoring rate to zero, the difference in PBIR no longer depends on $\lambda_{1diff}$, as presented in figure 4.3.10 and in appendix A.6. This finding reflects a limitation of our simulation model. By imposing at the beginning whether an individual is a responder or not, when a patient is censored by an administrative censoring, it does not leave the possibility that he/she changes status. Whereas in reality, it could happen otherwise. For example, let take patient A of figure 4.3.11. At the end of the study with follow-up of 7 months, no response time or progression is observed, so individual A will be censored. If our simulation study had defined him as a non-responder, he would have remained in this group, whereas perhaps if the end of the study was scheduled further out (e.g. follow-up of 20 months), he might have finally responded to the treatment.
Figure 4.3.10 also illustrated the influence of $\tau$. With a longer follow-up time, the value of $\tau$ is larger and therefore the AUC will be larger, as will the difference in PBIR.

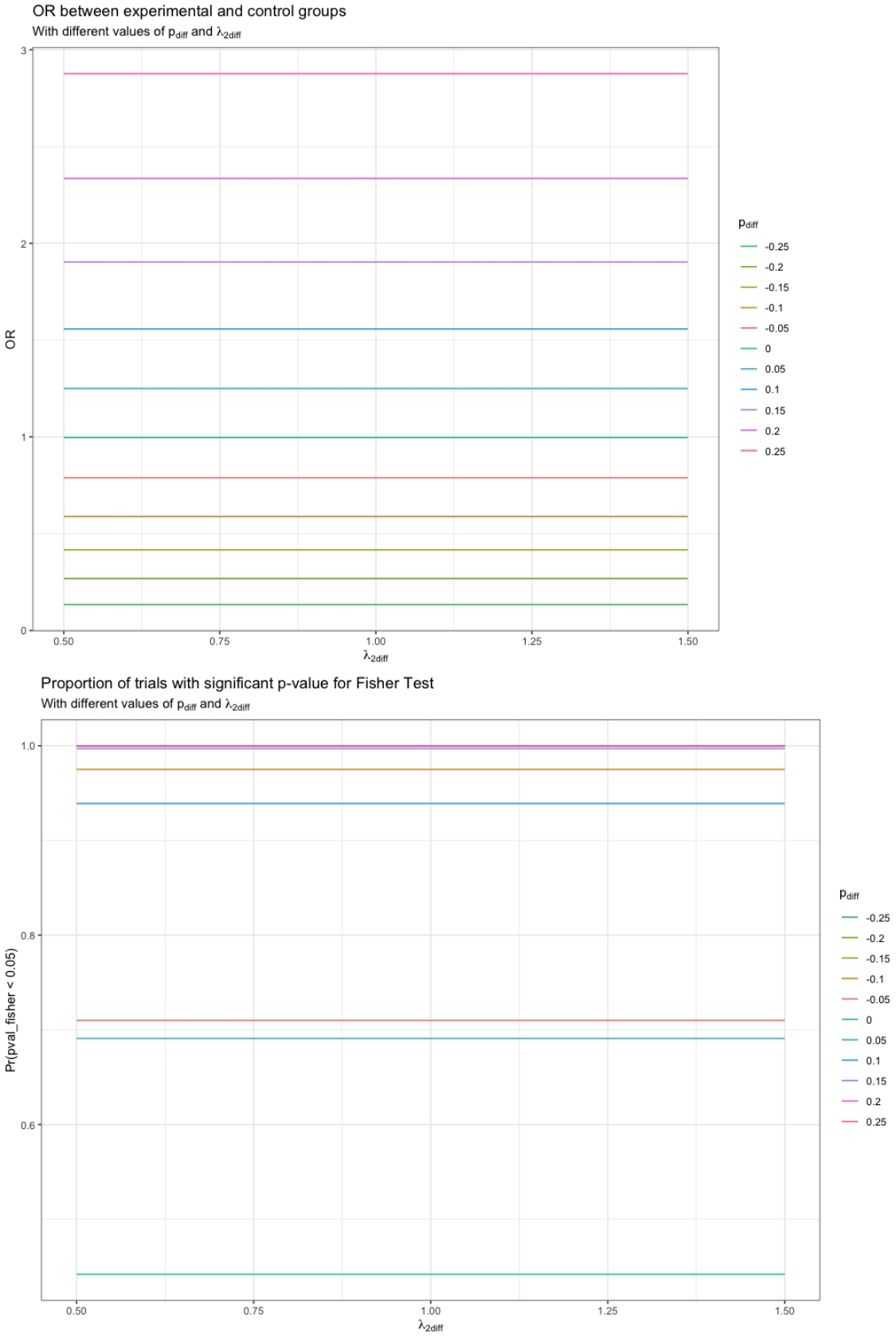


Figure 4.3.8: Simulation results for Fisher Test

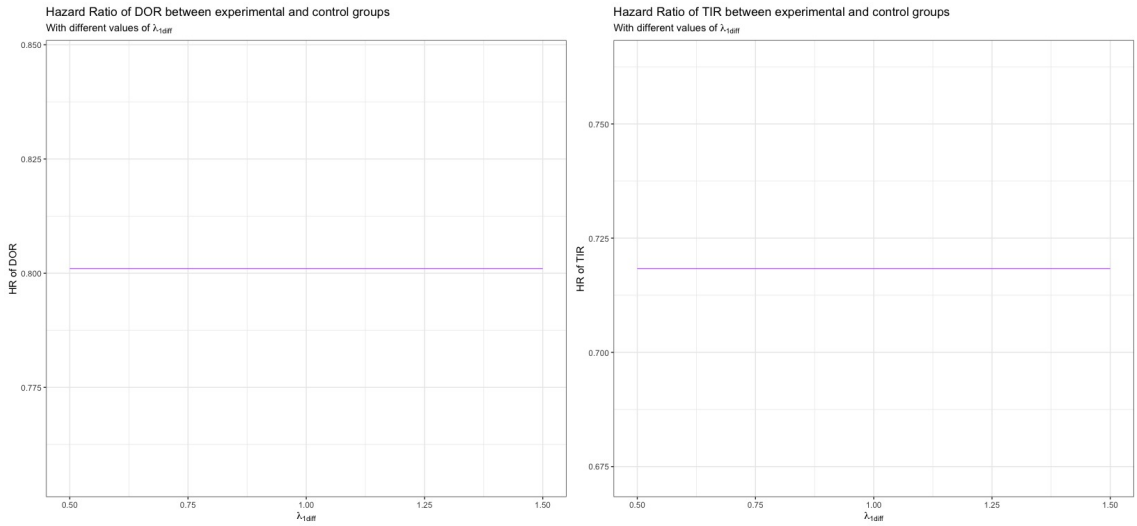


Figure 4.3.9: Simulation results for DOR and TIR with variations of $\lambda_{1diff}$

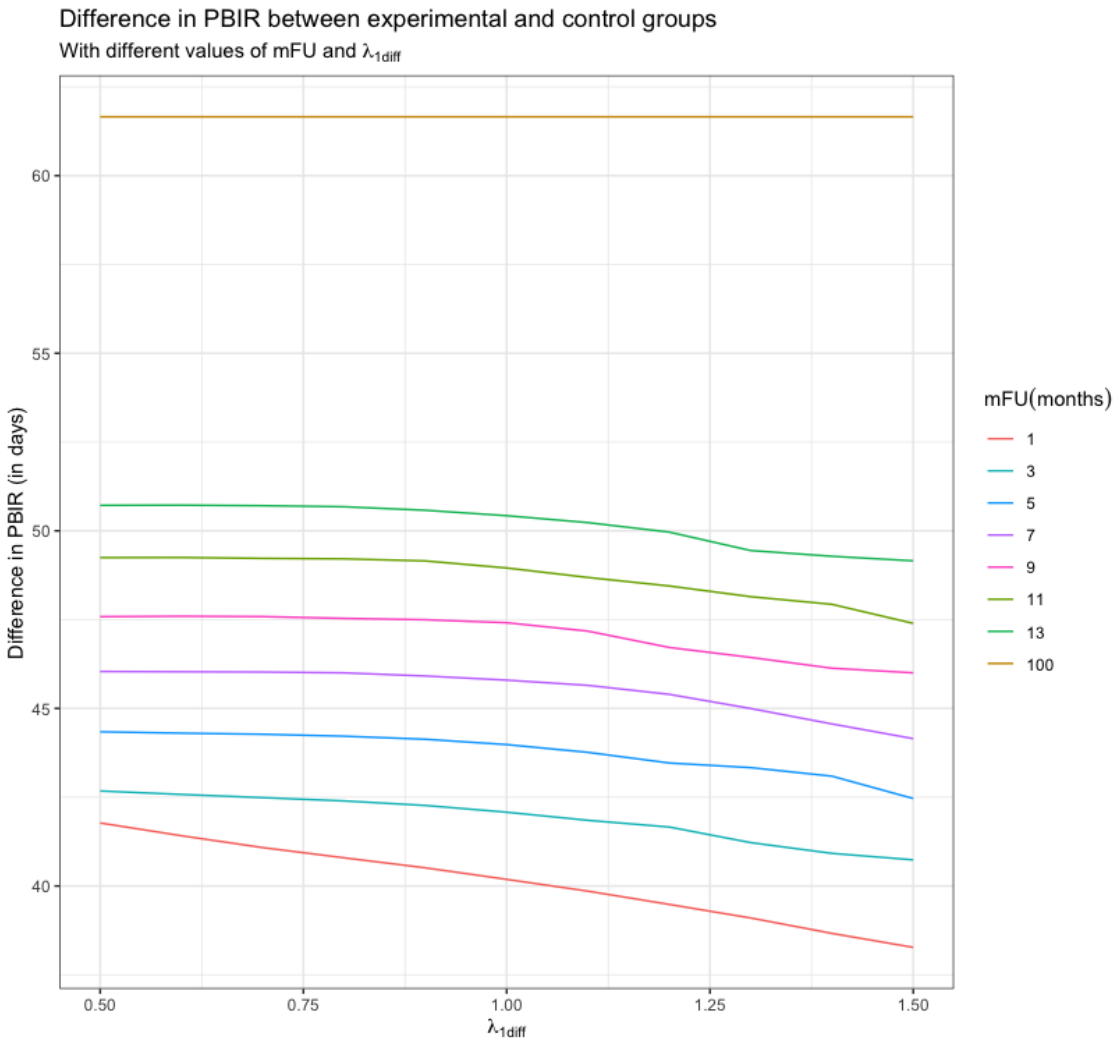


Figure 4.3.10: Simulation results for difference in PBIR with variations of $\lambda_{1diff}$

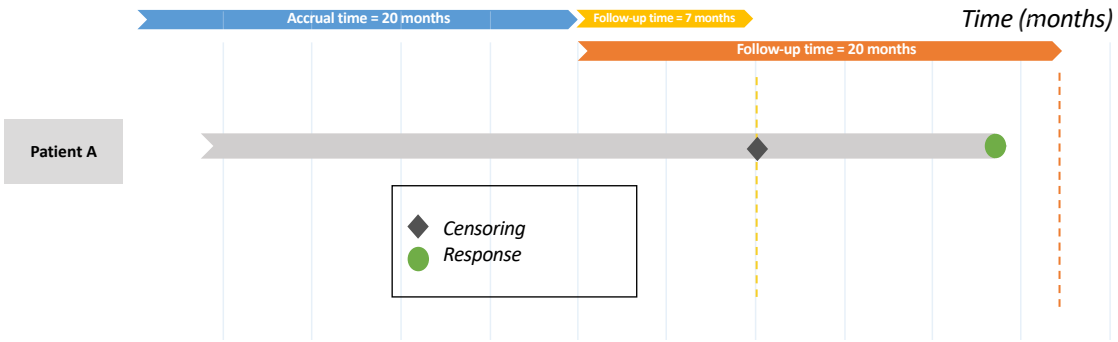


Figure 4.3.11: Status situations according to censoring

# Chapter 5

# Illustration in a clinical trial of head and neck cancer.

For this thesis, we have access to data in an oncology clinical trial from Project Data Sphere ® [1]. The open-label phase III randomized trial assesses the efficacy and safety of the combination of an experimental treatment and a control treatment (E+C) compared to the control treatment (C) for patients with recurrent or metastasis Squamous Cell Carcinoma of Head and Neck (SCCHN) [31]. The purpose of this illustration is first to analyze the trial through the traditional outcomes described by the FDA in Chapter 2. Then, the three methods of duration of response are applied and their results compared.

SCCHN is a group of cancer arising from mucosal surfaces in the oral cavity, pharynx and larynx and represents the most common malignancies in the head and neck and the sixth most common cancer worldwide with 450 000 deaths in 2018 [11]. The major etiologic factors of SCCHN were historically smoking and alcohol use but over the past decade, the Human Papilloma Virus (HPV) has been detected as a pathogen causing a distinct group of SCCHN, in particular Oropharyngeal Squamous Cell Carcinoma (SCC) [12]. Current standards of treatment for management of SCCHN are surgery, radiation and chemotherapy in different combinations, depending on the stage and on the localisation of primary tumor [15]. This study aims to investigate the effect of a new treatment combined with classical treatment (E+C) compared to the classical treatment (C).

Firstly, efficacy of the combination E+C was assessed by multiple endpoints : the primary endpoint concerned Overall Survival with censoring occurring if the patients are not dead at the time of the analysis with a censoring date corresponding to the last date they are known to be alive. Then, the secondary endpoints are Progression-Free Survival, Objective Response Rate and Time To Response (TTR) as the time from randomization to first confirmed objective response, and safety.

The available data counts 520 patients allocated in the two groups with 260 patients in each group and baseline characteristics of patients are presented in table 5.0.1. These char-

[1] This presentation is based on research using information obtained from data.projectdatasphere.org, which is maintained by Project Data Sphere. Neither Project Data Sphere nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this presentation.

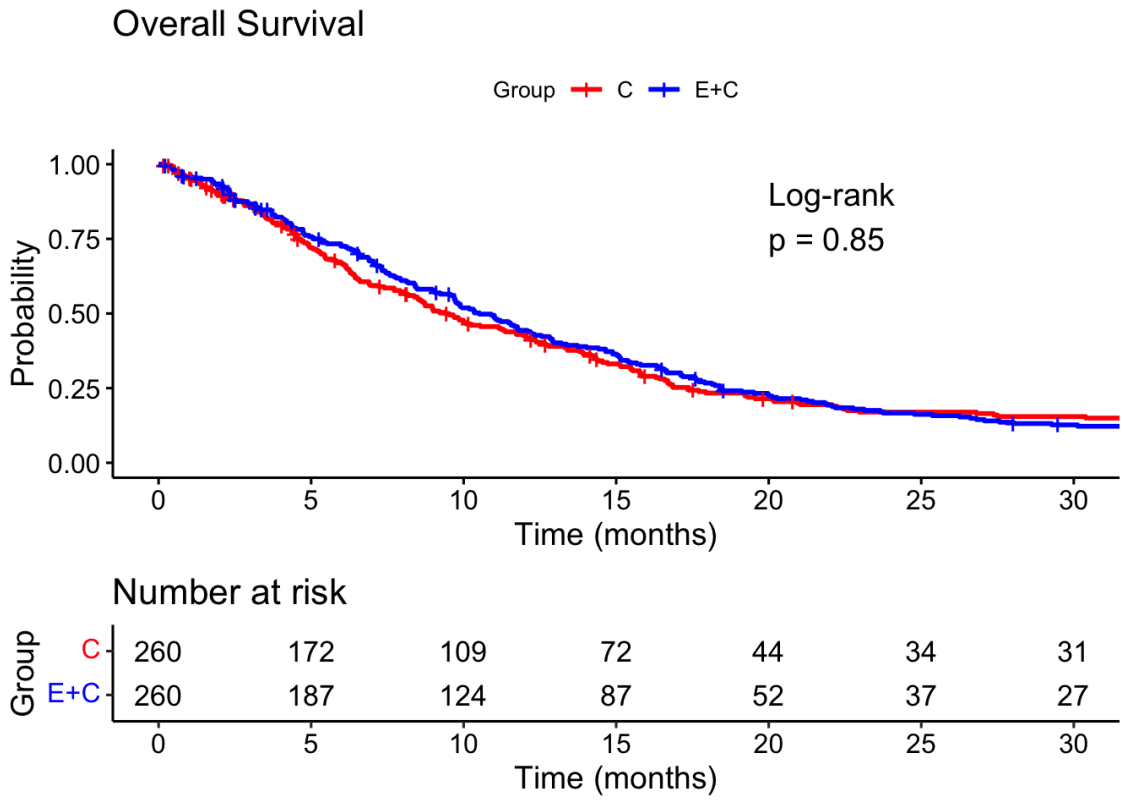


Figure 5.0.1: Illustration results - Kaplan-Meier curves by group for Overall Survival

acteristics seem to be distributed in the same proportion in both groups. Median OS is 10.51 months (95% CI [9.66 ; 12.22]) for E+C arm and 9.40 months (95% CI [8.38 ; 11.73]) for C arm (figure 5.0.1). The estimated HR is equal to 0.9821 (95% CI [0.8118 ; 1.188]) and the p-value of the log-rank test equals to 0.85, meaning a non-significant difference between groups for the primary endpoint.

For secondary endpoints, median PFS in the E+C group is 5.78 months (95% CI [5.49 ; 6.74]) and 4.63 months (95% CI [4.11 ; 5.52]) in the C group (figure 5.0.3) with an HR of 0.8561 (95% CI [0.7119 ; 1.029]). Summary results for PFS and OS can be found in table 5.0.2. Proportion of patients with a response (complete or partial) is significantly higher in the E+C group with respectively 22% and 31% responders in the E+C group and in the C group. The OR equals to 1.55 (95% CI [1.04 ; 2.30] ; p-value=0.0294) but median TTR are close in two groups with a median of about 6 weeks for both groups (table 5.0.3) . Empirical cumulative distribution function for TTR can be observed in figure 5.0.2. This distribution (similar for both groups) does not seem to follow a known distribution.

Secondly, the three methods for duration of response are used for this study and results are presented in figure 5.0.4. In the DOR method, we observe a longer Duration Of Response for the C arm. The corresponding HR is equal to 1.146 (95% CI [0.7907 ; 1.66]) representing a non-significant increase of the hazard of 14.6% in the E+C arm. Furthermore, the log-rank test presents a p-value of 0.5, meaning a non-significant difference between arms.

| | C group (n=260) | E+C group (n=260) |
|---|---|---|
| ***Sex*** | | |
| Female | 32 (12.31%) | 33 (12.69%) |
| Male | 228 (87.69%) | 227 (87.31%) |
| ***Age Category*** | | |
| <65 years | 207 (79.61%) | 220 (84.61%) |
| [65 ; 75 years) | 49 (18.85%) | 36 (13.85%) |
| ≥ 75 years | 4 (1.54%) | 4 (1.54%) |
| ***Race*** | | |
| Asian | 26 (10%) | 22 (8.46%) |
| Black or African American | 0 (0.00%) | 1 (0.38%) |
| Hispanic or Latino | 1 (0.38%) | 0 (0.00%) |
| White or Caucasian | 230 (88.46%) | 236 (90.77%) |
| Other | 1 (0.38%) | 1 (0.38%) |
| ***Primary Tumor Diagnosis*** | | |
| Hypopharynx | 32 (12.31%) | 43 (16.54%) |
| Larynx | 75 (28.85%) | 82 (31.54%) |
| Oral Cavity | 79 (30.38%) | 63 (24.23%) |
| Oropharynx | 74 (28.46%) | 72 (27.69%) |
| ***Primary Tumour Histological Type*** | | |
| Well differentiated | 54 (20.77%) | 53 (20.38%) |
| Moderately differentiated | 89 (34.23%) | 100 (38.46%) |
| Poorly differentiated | 53 (20.38%) | 43 (16.54%) |
| Undifferentiated | 5 (1.92%) | 5 (1.92%) |
| Not otherwise specified/unknown | 59 (22.69%) | 59 (22.69%) |
| ***Involountary Weight Loss Percentage*** | | |
| 0-5% | 20 (7.69%) | 17 (6.54%) |
| >5% | 54 (20.77%) | 46 (17.69%) |
| ***ECOG Performance Status*** | | |
| 0 : Fully active | 80 (30.77%) | 70 (26.92%) |
| 1 : Symptoms but ambulatory | 176 (67.69%) | 188 (72.31%) |
| 2 : In bed less than 50% of the time | 4 (1.54%) | 2 (0.77%) |
| ***Previous treatment (chemotherapy, radiotherapy or/and surgery)*** | | |
| Yes | 239 (92%) | 247 (95%) |
| No | 21 (8%) | 13 (5%) |

Data are expressed as n (%).

Data not available for all subjects. Missing data : Race, Involountary Weight Loss Percentage.
Abbreviation : ECOG Eastern Cooperative Oncology Group

Table 5.0.1: Illustration results - Baseline characteristics

| | C group (n=260) | E+C group (n=260) | Hazard Ratio |
|---|---|---|---|
| ***OS (months)*** | 9.4 [8.38 ; 11.73] | 10.51 [9.66 ; 12.22] | 0.9821 [0.8118 ; 1.188] |
| ***PFS (months)*** | 4.63 [4.11 ; 5.52] | 5.78 [5.49 ; 6.74] | 0.8561 [0.7119 ; 1.029] |
| Data are expressed as median [95% CI] and estimate [95% CI] for Hazard Ratio. | | | |

Table 5.0.2: Illustration results - Results for PFS and OS

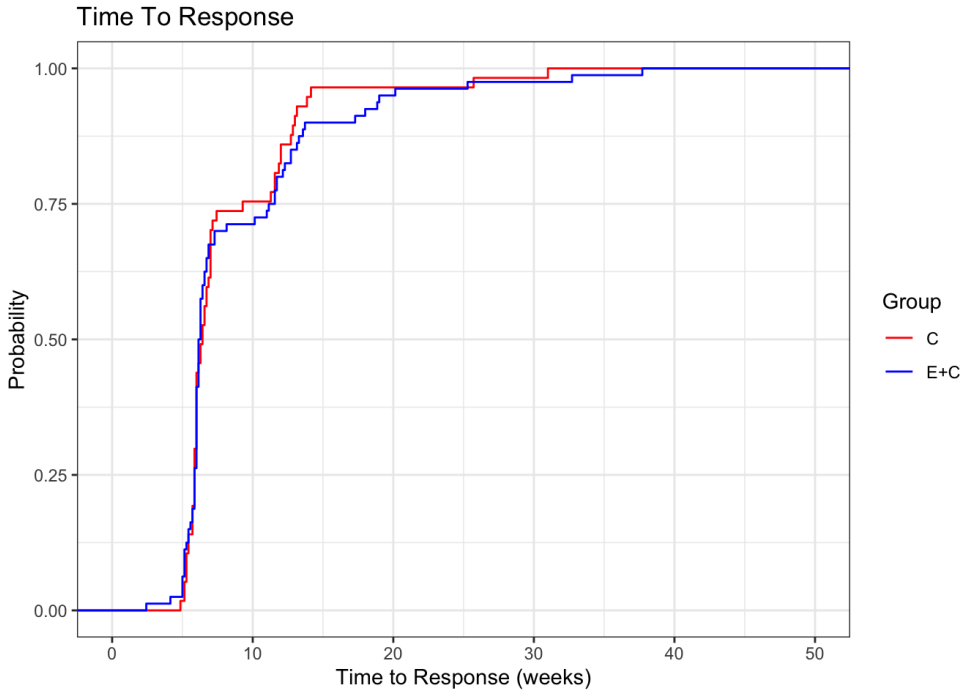


Figure 5.0.2: Illustration results - Empirical Cumulative Distribution Function for Time To Response

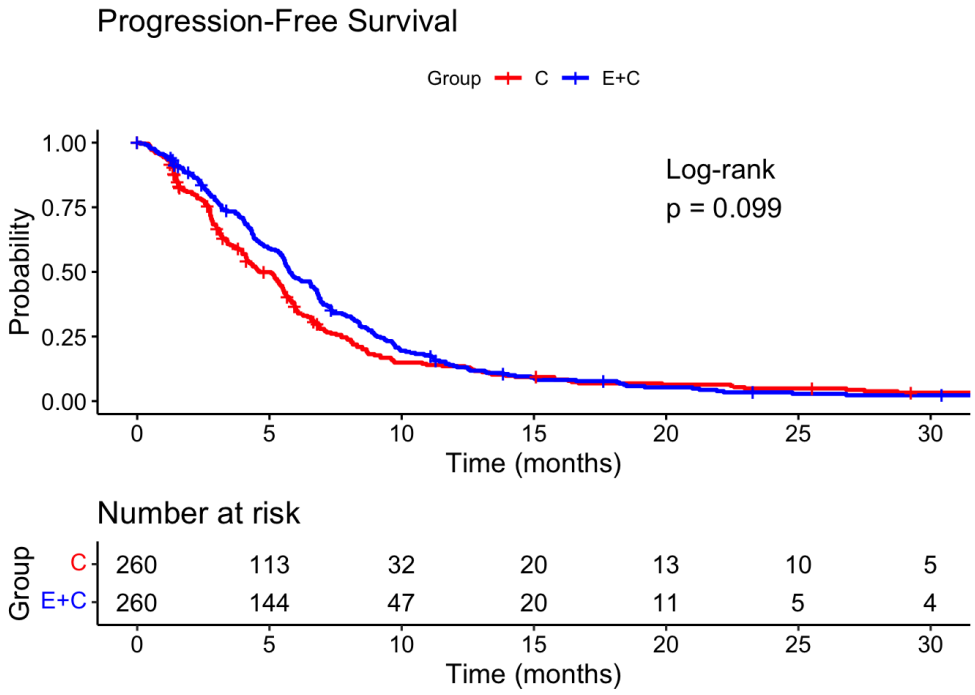


Figure 5.0.3: Illustration results - Kaplan-Meier curves by group for Progression-Free Survival

| | **C group (n=260)** | **E+C group (n=260)** |
|---|---|---|
| ***Objective Response Rate*** | | |
| Complete response | 7 (2.69%) | 9 (3.46%) |
| Partial response | 51 (19.62%) | 71 (27.31%) |
| Stable disease | 133 (51.15%) | 134 (51.54%) |
| Disease progression | 31 (11.92%) | 17 (6.54%) |
| Unevaluable | 2 (0.77%) | 4 (1.54%) |
| Not done | 36 (13.85%) | 25 (9.62%) |
| ***Time To Response (weeks)**** | 6.43 (5.86 - 10.79) | 6.21 (5.86 - 11.25) |

Data are expressed as n(%) or median(IQR).
*Included only patients with a complete response or a partial response.

Table 5.0.3: Illustration results - Results for the response

| | *Estimate* | *Confidence interval* | *Associated p-value* |
|---|---|---|---|
| **OS** | HR = 0.9821 | [0.8118 ; 1.188] | 0.85 |
| **PFS** | HR = 0.8561 | [0.7119 ; 1.029] | 0.10 |
| **DOR** | HR = 1.146 | [0.7907 ; 1.66] | 0.47 |
| **TIR** | HR = 0.8795 | [0.7379 ; 1.048] | 0.15 |
| **PBIR** (days) | MDOR = 12.2602 | [-24.6067 ; 49.127] | 0.5145 |

Table 5.0.4: Illustration results - Summary of results

This conventional analysis is assessed only for responders, so the selection is based on a post-randomization criterion. Thus, the comparison of DOR is not fair because response rate is higher in the E+C group, compared to C group. Indeed, the DOR is longer for C group but the response rate is higher for the E+C group, which leads the clinician to a difficulty to conclude which treatment is the best.

Time In Response includes all the patients with an artificial event at day zero for non-responders. Patients remaining after the fall at time 0 correspond to the proportion of responders in each arm, i.e. 22% in the C group and 31% in the E+C group. In this second method, the difference between the two groups (except during the fall at time 0) is no longer as important as in the DOR method. The corresponding HR is 0.879 (95% CI [0.7348 ; 1.051]) so being in the E+C group reduces the hazard of 12.1%. The log-rank test does not show a significant difference between arms with a p-value of 0.15.

Finally, the third method of Probability of Being In Response is tested. We observe that the PBIR curve of the E+C group looks higher until month 20, but the difference seems the most important before month 10. This can be confirmed by the curve of the difference in PBIR between the two treatment arms in time. Looking at the confidence intervals, this difference does not appear to be significant. Moreover, the estimation of the mean duration of response provides a global summary measure for the treatment difference and it appears that the mean duration of response in the time window [0 ; 39.33] between the two groups is 12.26 days (95% CI [-24.61 ; 49.127]) with a corresponding p-value of Wald-test of 0.5145, showing a non-significant difference at the 5% significance level. Note that the 39.33-month time window corresponds to the maximum study follow-up.

As a conclusion, we can compare results of traditional endpoints and from the three methods for duration of response in table 5.0.4. The DOR estimate shows a higher risk of having the event of interest in the E+C group, i.e. the duration of response is shorter in this group, whereas the other endpoints studied are in favor of the experimental group with (1) a survival time, (2) a time to progression, (3) a time in response and (4) a mean duration of response, longer in the E+C group compared to the C group. Nevertheless, none of these differences is significant between groups with an associated p-value always larger than 0.05.

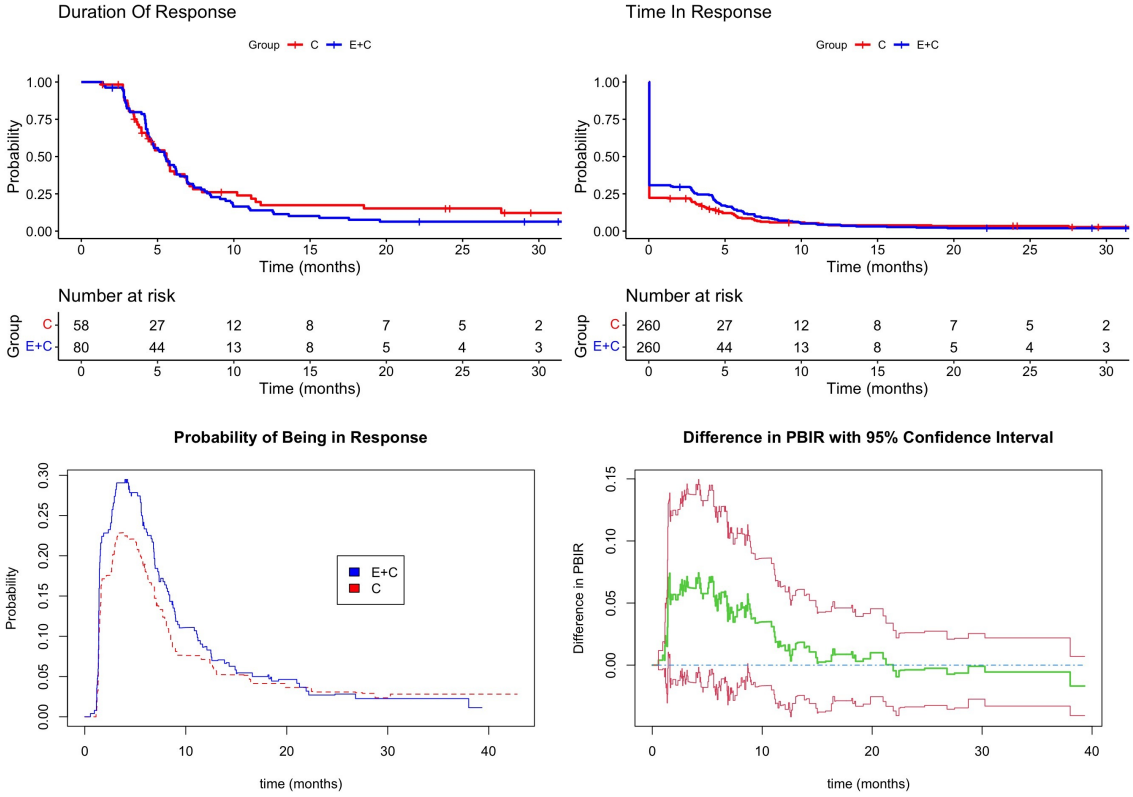


Figure 5.0.4: Illustration results - Resulting plots for three methods

# Chapter 6

# Discussion.

The main objective of this master thesis was to review, apply and compare different methods for assessing duration of response in oncology studies. On the one hand, we compared these three methods on a simulation study by observing the impact of two parameters, the duration of response and the response rate, on each of the methods. On the other hand, we were able to perform this comparison by applying them to data from a real study.

To begin, the Duration Of Response (DOR), assessed as the time from response to progression or death, whichever comes first, showed an influence only of the duration of response parameter. The shorter is the duration of response in the experimental arm (relative to the control arm), the higher is the Hazards Ratio, showing a greater risk of progression or death in the experimental arm. Nevertheless, during the analysis on real data, we were confronted with the main problem of the Duration Of Response method, with contradictory results when one group has a longer duration of response but fewer responders than the other treatment arm, which makes it difficult to establish which treatment is the best, based on the response to treatment.

Next, the Time In Response (TIR) method allows to include all the patients by adding artificial event to non-responders. The TIR showed an influence of the response rate and the duration of response. For the same duration of response, a higher response rate in the experimental arm shows a lower HR and for the same response rate, a longer duration of response shows a lower HR as well. It should be noted that there does not seem to be any interaction between these parameters, i.e. the duration of response influences the HR in the same way, regardless of the response rate and vice-versa.

Finally, the method introduced by Huang and al., the Probability of Being In Response (PBIR) curve, evaluates the difference in mean duration of response between groups. This difference is influenced by both response rate and duration of response. Unlike the TIR method, the influence of duration of response is stronger when the response rate of the experimental arm is large. When the duration of response is longer in the experimental arm, the difference in PBIR is larger if there are more responders in this arm. Moreover, the conclusions drawn by the Wald statistical test associated with the PBIR method seem to give similar results to those drawn in the TIR method. Thanks to the application on real data, we can observe that implementing this method is feasible and allows an easier interpretation of the results.

When interpreting the results of this master thesis, it is important to consider various limitations encountered. Even if the chosen simulation model allows to evaluate the effect of duration of response and response rate on the different methods, several limitations occur. Indeed, the model to generate event times and censoring could have been more general (e.g. Weibull distributions can be used or the status of responder or non-responder may not be imposed on the first step of the model as done here) and the simulations could have been enlarged to consider the impact of other parameters as the accrual time and the follow-up time or the value of $\tau$, representing the chosen time window.

In this work, we consider a non-parametric estimator of the PBIR function, which is the difference between two Kaplan-Meier estimates. Tsai and al.(2017) [29] have developed two other estimators for this function. These estimators were not discussed here because their clinical interpretation is less intuitive but it is important to note their existence and to compare these estimators in further work.

No study to my knowledge has addressed the issue of the required sample size based on the PBIR method. However, Royston and al.(2013) [24] investigated the sample size needed to assess the difference in the restricted mean survival time of two treatment groups. Based on his work, it would be interesting to study the sample size and to evaluate the different parameters influencing this size.

In conclusion, this master thesis studied three methods for comparing the duration of response between two treatments in oncology. The Duration Of Response method based on a post-randomization criterion with only a sub-group of patients, can lead to confusing HR and result of a log-rank test. It is therefore preferable to limit this method to a descriptive analysis of responders.
Although we were able to highlight the possibility of calculating an HR and performing a log-rank test using the Time In Response method, the interpretation of these results using this method remains difficult due to the presence of artificial events added to the non-responders. Therefore, the Probability of Being In Response method proposed by Huang and al. seems to be an interesting alternative to estimate the duration of response in a clinical trial while taking into account the response rate. By combining these two pieces of information and taking into account all patients for the whole duration of the study, this estimation allows an intuitive interpretation by overcoming the different problems encountered in the Duration Of Response and Time In Response methods.

# Appendix

## A.1 Tables of the results of the simulation study for DOR Method.

| $p_{diff} \backslash \lambda_{2diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.25** | 0.49 | 0.59 | 0.69 | 0.80 | 0.90 | 1.00 | 1.11 | 1.21 | 1.32 | 1.42 | 1.53 |
| **-0.2** | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 | 1.11 | 1.21 | 1.31 | 1.42 | 1.52 |
| **-0.15** | 0.49 | 0.59 | 0.69 | 0.80 | 0.90 | 1.00 | 1.10 | 1.20 | 1.30 | 1.40 | 1.50 |
| **-0.1** | 0.49 | 0.59 | 0.69 | 0.79 | 0.89 | 0.99 | 1.09 | 1.19 | 1.29 | 1.39 | 1.49 |
| **-0.05** | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.01 | 1.11 | 1.21 | 1.31 | 1.41 | 1.51 |
| **0** | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 | 1.10 | 1.20 | 1.30 | 1.40 | 1.51 |
| **0.05** | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 | 1.10 | 1.20 | 1.30 | 1.40 | 1.50 |
| **0.1** | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.01 | 1.11 | 1.21 | 1.31 | 1.41 | 1.51 |
| **0.15** | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 | 1.10 | 1.21 | 1.31 | 1.41 | 1.51 |
| **0.2** | 0.50 | 0.61 | 0.71 | 0.81 | 0.91 | 1.01 | 1.11 | 1.21 | 1.31 | 1.41 | 1.52 |
| **0.25** | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 | 1.10 | 1.20 | 1.30 | 1.40 | 1.50 |

Table A.1.1: Hazard ratio of DOR

| $p_{diff} \backslash \lambda_{2diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.25** | 0.466 | 0.285 | 0.177 | 0.101 | 0.069 | 0.054 | 0.072 | 0.098 | 0.146 | 0.215 | 0.290 |
| **-0.2** | 0.703 | 0.470 | 0.271 | 0.126 | 0.053 | 0.055 | 0.089 | 0.141 | 0.205 | 0.296 | 0.414 |
| **-0.15** | 0.816 | 0.590 | 0.365 | 0.175 | 0.091 | 0.051 | 0.078 | 0.154 | 0.239 | 0.347 | 0.479 |
| **-0.1** | 0.893 | 0.670 | 0.405 | 0.216 | 0.080 | 0.057 | 0.086 | 0.143 | 0.254 | 0.406 | 0.551 |
| **-0.05** | 0.919 | 0.710 | 0.427 | 0.213 | 0.089 | 0.059 | 0.100 | 0.193 | 0.310 | 0.461 | 0.617 |
| **0** | 0.944 | 0.760 | 0.487 | 0.228 | 0.093 | 0.063 | 0.075 | 0.172 | 0.315 | 0.481 | 0.649 |
| **0.05** | 0.954 | 0.797 | 0.534 | 0.246 | 0.094 | 0.056 | 0.086 | 0.188 | 0.346 | 0.510 | 0.674 |
| **0.1** | 0.970 | 0.827 | 0.529 | 0.237 | 0.087 | 0.063 | 0.095 | 0.200 | 0.374 | 0.548 | 0.698 |
| **0.15** | 0.984 | 0.848 | 0.538 | 0.256 | 0.097 | 0.051 | 0.093 | 0.194 | 0.387 | 0.569 | 0.717 |
| **0.2** | 0.983 | 0.830 | 0.560 | 0.252 | 0.089 | 0.038 | 0.084 | 0.223 | 0.389 | 0.582 | 0.752 |
| **0.25** | 0.985 | 0.863 | 0.611 | 0.289 | 0.119 | 0.056 | 0.089 | 0.197 | 0.380 | 0.581 | 0.749 |

Table A.1.2: Proportion of trials with p-value $\leq 0.05$

## A.2 Tables of the results of the simulation study for TIR Method.

| $p_{diff} \backslash \lambda_{2diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.25** | 1.70 | 1.73 | 1.75 | 1.77 | 1.79 | 1.81 | 1.83 | 1.85 | 1.86 | 1.88 | 1.89 |
| **-0.2** | 1.42 | 1.45 | 1.47 | 1.50 | 1.52 | 1.55 | 1.57 | 1.59 | 1.61 | 1.63 | 1.64 |
| **-0.15** | 1.23 | 1.26 | 1.28 | 1.31 | 1.34 | 1.36 | 1.38 | 1.41 | 1.43 | 1.45 | 1.47 |
| **-0.1** | 1.08 | 1.11 | 1.13 | 1.16 | 1.19 | 1.21 | 1.24 | 1.26 | 1.28 | 1.30 | 1.32 |
| **-0.05** | 0.95 | 0.98 | 1.01 | 1.04 | 1.07 | 1.10 | 1.12 | 1.14 | 1.17 | 1.19 | 1.21 |
| **0** | 0.86 | 0.89 | 0.92 | 0.95 | 0.98 | 1.00 | 1.03 | 1.05 | 1.07 | 1.09 | 1.12 |
| **0.05** | 0.77 | 0.80 | 0.83 | 0.86 | 0.89 | 0.92 | 0.94 | 0.97 | 0.99 | 1.01 | 1.03 |
| **0.1** | 0.69 | 0.72 | 0.76 | 0.78 | 0.81 | 0.84 | 0.86 | 0.89 | 0.91 | 0.93 | 0.95 |
| **0.15** | 0.63 | 0.66 | 0.69 | 0.72 | 0.75 | 0.77 | 0.80 | 0.82 | 0.85 | 0.87 | 0.89 |
| **0.2** | 0.57 | 0.60 | 0.63 | 0.66 | 0.69 | 0.71 | 0.74 | 0.76 | 0.79 | 0.81 | 0.83 |
| **0.25** | 0.51 | 0.54 | 0.57 | 0.60 | 0.63 | 0.65 | 0.68 | 0.70 | 0.73 | 0.75 | 0.77 |

Table A.2.1: Hazard ratio of TIR

| $p_{diff} \backslash \lambda_{2diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.25** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **-0.2** | 0.980 | 0.992 | 0.993 | 0.996 | 0.997 | 0.997 | 0.998 | 0.999 | 0.999 | 0.999 | 1.000 |
| **-0.15** | 0.623 | 0.726 | 0.808 | 0.873 | 0.924 | 0.939 | 0.960 | 0.974 | 0.985 | 0.985 | 0.991 |
| **-0.1** | 0.112 | 0.196 | 0.264 | 0.362 | 0.464 | 0.570 | 0.665 | 0.747 | 0.802 | 0.846 | 0.895 |
| **-0.05** | 0.072 | 0.039 | 0.040 | 0.058 | 0.116 | 0.157 | 0.221 | 0.303 | 0.386 | 0.474 | 0.561 |
| **0** | 0.385 | 0.239 | 0.129 | 0.070 | 0.049 | 0.047 | 0.056 | 0.076 | 0.118 | 0.168 | 0.227 |
| **0.05** | 0.792 | 0.666 | 0.510 | 0.363 | 0.236 | 0.151 | 0.104 | 0.068 | 0.060 | 0.051 | 0.069 |
| **0.1** | 0.977 | 0.938 | 0.867 | 0.743 | 0.624 | 0.473 | 0.347 | 0.241 | 0.166 | 0.114 | 0.082 |
| **0.15** | 0.998 | 0.992 | 0.977 | 0.952 | 0.896 | 0.808 | 0.702 | 0.564 | 0.451 | 0.332 | 0.228 |
| **0.2** | 1.000 | 1.000 | 0.999 | 0.997 | 0.986 | 0.954 | 0.911 | 0.845 | 0.758 | 0.643 | 0.524 |
| **0.25** | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 0.996 | 0.987 | 0.972 | 0.939 | 0.874 | 0.816 |

Table A.2.2: Proportion of trials with p-value $\leq 0.05$

## A.3 Tables of the results of the simulation study for PBIR Method.

| $p_{diff} \backslash \lambda_{2diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.25** | -43.37 | -44.14 | -44.83 | -45.45 | -46.02 | -46.56 | -47.06 | -47.50 | -47.90 | -48.29 | -48.64 |
| **-0.2** | -31.38 | -32.86 | -34.19 | -35.42 | -36.54 | -37.55 | -38.49 | -39.34 | -40.14 | -40.86 | -41.52 |
| **-0.15** | -18.96 | -21.19 | -23.22 | -25.11 | -26.80 | -28.36 | -29.76 | -31.03 | -32.21 | -33.30 | -34.29 |
| **-0.1** | -6.00 | -9.00 | -11.8 | -14.35 | -16.65 | -18.73 | -20.64 | -22.37 | -23.95 | -25.39 | -26.71 |
| **-0.05** | 7.23 | 3.30 | -0.32 | -3.62 | -6.60 | -9.30 | -11.75 | -13.98 | -16.01 | -17.85 | -19.55 |
| **0** | 19.80 | 15.06 | 10.69 | 6.65 | 3.04 | -0.25 | -3.21 | -5.89 | -8.34 | -10.56 | -12.61 |
| **0.05** | 33.56 | 27.81 | 22.48 | 17.66 | 13.35 | 9.42 | 5.88 | 2.66 | -0.24 | -2.87 | -5.26 |
| **0.1** | 48.32 | 41.34 | 35.05 | 29.28 | 24.14 | 19.51 | 15.35 | 11.59 | 8.19 | 5.12 | 2.32 |
| **0.15** | 61.91 | 54.07 | 46.77 | 40.36 | 34.52 | 29.20 | 24.44 | 20.15 | 16.27 | 12.77 | 9.60 |
| **0.2** | 76.24 | 67.31 | 59.10 | 51.68 | 45.01 | 39.00 | 33.61 | 28.79 | 24.43 | 20.48 | 16.87 |
| **0.25** | 92.02 | 81.89 | 72.56 | 64.10 | 56.65 | 49.97 | 43.98 | 38.57 | 33.67 | 29.24 | 25.22 |

Table A.3.1: Difference in PBIR

| $p_{diff} \backslash \lambda_{2diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.25** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **-0.2** | 0.956 | 0.967 | 0.986 | 0.994 | 0.997 | 0.998 | 0.998 | 0.998 | 0.999 | 0.999 | 0.999 |
| **-0.15** | 0.522 | 0.635 | 0.731 | 0.820 | 0.880 | 0.919 | 0.947 | 0.963 | 0.982 | 0.987 | 0.992 |
| **-0.1** | 0.104 | 0.170 | 0.247 | 0.343 | 0.425 | 0.551 | 0.657 | 0.743 | 0.814 | 0.860 | 0.901 |
| **-0.05** | 0.101 | 0.066 | 0.053 | 0.060 | 0.110 | 0.166 | 0.232 | 0.326 | 0.445 | 0.544 | 0.650 |
| **0** | 0.457 | 0.302 | 0.181 | 0.094 | 0.071 | 0.054 | 0.066 | 0.109 | 0.158 | 0.231 | 0.312 |
| **0.05** | 0.811 | 0.690 | 0.538 | 0.391 | 0.251 | 0.168 | 0.106 | 0.071 | 0.068 | 0.081 | 0.107 |
| **0.1** | 0.979 | 0.935 | 0.868 | 0.752 | 0.612 | 0.461 | 0.335 | 0.215 | 0.146 | 0.096 | 0.071 |
| **0.15** | 0.998 | 0.994 | 0.972 | 0.943 | 0.884 | 0.782 | 0.649 | 0.522 | 0.365 | 0.254 | 0.179 |
| **0.2** | 1.000 | 1.000 | 0.999 | 0.993 | 0.976 | 0.941 | 0.873 | 0.790 | 0.665 | 0.531 | 0.429 |
| **0.25** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.993 | 0.980 | 0.958 | 0.906 | 0.816 | 0.715 |

Table A.3.2: Proportion of trials with p-value $\leq 0.05$

## A.4 Table of the results of the simulation study for censoring rate.

| $p_{diff} \backslash \lambda_{2diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.25** | 9.38 | 9.20 | 9.06 | 8.94 | 8.83 | 8.74 | 8.66 | 8.60 | 8.54 | 8.49 | 8.45 |
| **-0.2** | 10.28 | 9.96 | 9.69 | 9.46 | 9.27 | 9.11 | 8.96 | 8.85 | 8.74 | 8.65 | 8.57 |
| **-0.15** | 11.29 | 10.79 | 10.39 | 10.04 | 9.76 | 9.52 | 9.31 | 9.14 | 8.99 | 8.86 | 8.74 |
| **-0.1** | 12.13 | 11.51 | 10.97 | 10.53 | 10.16 | 9.84 | 9.56 | 9.33 | 9.14 | 8.97 | 8.82 |
| **-0.05** | 13.21 | 12.42 | 11.73 | 11.18 | 10.71 | 10.31 | 9.97 | 9.68 | 9.43 | 9.21 | 9.03 |
| **0** | 14.17 | 13.24 | 12.45 | 11.78 | 11.23 | 10.74 | 10.34 | 9.99 | 9.70 | 9.44 | 9.21 |
| **0.05** | 15.13 | 14.04 | 13.15 | 12.36 | 11.70 | 11.14 | 10.67 | 10.25 | 9.91 | 9.62 | 9.37 |
| **0.1** | 16.08 | 14.83 | 13.79 | 12.90 | 12.16 | 11.52 | 10.99 | 10.54 | 10.15 | 9.81 | 9.52 |
| **0.15** | 17.03 | 15.65 | 14.47 | 13.48 | 12.65 | 11.93 | 11.33 | 10.80 | 10.37 | 9.99 | 9.66 |
| **0.2** | 17.98 | 16.41 | 15.12 | 14.05 | 13.13 | 12.33 | 11.67 | 11.10 | 10.62 | 10.19 | 9.81 |
| **0.25** | 19.00 | 17.29 | 15.84 | 14.63 | 13.63 | 12.75 | 12.02 | 11.37 | 10.82 | 10.35 | 9.95 |

Table A.4.1: Censoring rates (in %)

## A.5 Tables of the results of the simulation study for OR and for Fisher Test.

| $p_{diff} \backslash \lambda_{2diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.25** | 0.1337 | 0.1337 | 0.1337 | 0.1337 | 0.1337 | 0.1337 | 0.1337 | 0.1337 | 0.1337 | 0.1337 | 0.1337 |
| **-0.2** | 0.2681 | 0.2681 | 0.2681 | 0.2681 | 0.2681 | 0.2681 | 0.2681 | 0.2681 | 0.2681 | 0.2681 | 0.2681 |
| **-0.15** | 0.4166 | 0.4166 | 0.4166 | 0.4166 | 0.4166 | 0.4166 | 0.4166 | 0.4166 | 0.4166 | 0.4166 | 0.4166 |
| **-0.1** | 0.5884 | 0.5884 | 0.5884 | 0.5884 | 0.5884 | 0.5884 | 0.5884 | 0.5884 | 0.5884 | 0.5884 | 0.5884 |
| **-0.05** | 0.7888 | 0.7888 | 0.7888 | 0.7888 | 0.7888 | 0.7888 | 0.7888 | 0.7888 | 0.7888 | 0.7888 | 0.7888 |
| **0** | 0.9962 | 0.9962 | 0.9962 | 0.9962 | 0.9962 | 0.9962 | 0.9962 | 0.9962 | 0.9962 | 0.9962 | 0.9962 |
| **0.05** | 1.2497 | 1.2497 | 1.2497 | 1.2497 | 1.2497 | 1.2497 | 1.2497 | 1.2497 | 1.2497 | 1.2497 | 1.2497 |
| **0.1** | 1.5575 | 1.5575 | 1.5575 | 1.5575 | 1.5575 | 1.5575 | 1.5575 | 1.5575 | 1.5575 | 1.5575 | 1.5575 |
| **0.15** | 1.9028 | 1.9028 | 1.9028 | 1.9028 | 1.9028 | 1.9028 | 1.9028 | 1.9028 | 1.9028 | 1.9028 | 1.9028 |
| **0.2** | 2.3351 | 2.3351 | 2.3351 | 2.3351 | 2.3351 | 2.3351 | 2.3351 | 2.3351 | 2.3351 | 2.3351 | 2.3351 |
| **0.25** | 2.8763 | 2.8763 | 2.8763 | 2.8763 | 2.8763 | 2.8763 | 2.8763 | 2.8763 | 2.8763 | 2.8763 | 2.8763 |

Table A.5.1: OR

| $p_{diff} \backslash \lambda_{2diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **-0.25** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **-0.2** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **-0.15** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **-0.1** | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 | 0.975 |
| **-0.05** | 0.710 | 0.710 | 0.710 | 0.710 | 0.710 | 0.710 | 0.710 | 0.710 | 0.710 | 0.710 | 0.710 |
| **0** | 0.442 | 0.442 | 0.442 | 0.442 | 0.442 | 0.442 | 0.442 | 0.442 | 0.442 | 0.442 | 0.442 |
| **0.05** | 0.691 | 0.691 | 0.691 | 0.691 | 0.691 | 0.691 | 0.691 | 0.691 | 0.691 | 0.691 | 0.691 |
| **0.1** | 0.939 | 0.939 | 0.939 | 0.939 | 0.939 | 0.939 | 0.939 | 0.939 | 0.939 | 0.939 | 0.939 |
| **0.15** | 0.997 | 0.997 | 0.997 | 0.997 | 0.997 | 0.997 | 0.997 | 0.997 | 0.997 | 0.997 | 0.997 |
| **0.2** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **0.25** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table A.5.2: Proportion of trials with p-value $\leq 0.05$ from Fisher Test

## A.6 Table of the results of the simulation study for PBIR Method with variations of $\lambda_{1diff}$.

| $mFU \backslash \lambda_{1diff}$ | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** | **1.1** | **1.2** | **1.3** | **1.4** | **1.5** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 41.7759 | 41.4147 | 41.0816 | 40.7939 | 40.5117 | 40.1869 | 39.8551 | 39.4814 | 39.1007 | 38.6657 | 38.2722 |
| **3** | 42.6759 | 42.5783 | 42.4874 | 42.3951 | 42.2679 | 42.0777 | 41.8493 | 41.6605 | 41.2219 | 40.9187 | 40.7367 |
| **5** | 44.3419 | 44.3027 | 44.2727 | 44.2180 | 44.1320 | 43.9814 | 43.7631 | 43.4634 | 43.3329 | 43.0926 | 42.4634 |
| **7** | 46.0404 | 46.0321 | 46.0247 | 46.0013 | 45.9129 | 45.7959 | 45.6491 | 45.3966 | 44.9980 | 44.5631 | 44.1491 |
| **9** | 47.5833 | 47.5952 | 47.5855 | 47.5344 | 47.4968 | 47.4123 | 47.1763 | 46.7158 | 46.4367 | 46.1320 | 46.0021 |
| **11** | 49.2465 | 49.2494 | 49.2247 | 49.2113 | 49.1526 | 48.9543 | 48.6870 | 48.4471 | 48.1467 | 47.9313 | 47.3944 |
| **13** | 50.7182 | 50.7234 | 50.7079 | 50.6784 | 50.5793 | 50.4264 | 50.2314 | 49.9638 | 49.4453 | 49.2837 | 49.1558 |
| **100** | 61.6564 | 61.6564 | 61.6564 | 61.6564 | 61.6564 | 61.6564 | 61.6564 | 61.6564 | 61.6564 | 61.6564 | 61.6564 |

# Bibliography


[1] Abd ElHafeez S, D'Arrigo G, Leonardis D, Fusaro M, Tripepi G, Roumeliotis S. Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. Oxid Med Cell Longev. 2021 Nov 30;2021:1302811. doi: 10.1155/2021/1302811. PMID: 34887996; PMCID: PMC8651375.

[2] Brody, T. (2016). Chapter 13 - Oncology Endpoints : Overall Survival and Progression-Free Survival. In Clinical Trials (Second Edition) (p. 269 to 288). Academic Press. https://doi.org/10.1016/B978-0-12-804217-5.00013-8

[3] Delgado A, Guddati AK. Clinical endpoints in oncology - a primer. Am J Cancer Res. 2021 Apr 15;11(4):1121-1131. PMID: 33948349; PMCID: PMC8085844.

[4] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009 Jan;45(2):228-47. doi: 10.1016/j.ejca.2008.10.026. PMID: 19097774.

[5] Ellis S, Carroll KJ, Pemberton K. Analysis of duration of response in oncology trials. Contemporary clinical trials. 2008 7 1 ;29(4) :456 - 65. [PubMed : 18187370]

[6] EMA. Draft Guideline on the clinical evaluation of anticancer medicinal products in man. 2017.

[7] Clinical Trial Endpoints for Approval of Cancer Drugs and Biologics. December 2018. U.S. Food and Drug Administration. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics

[8] Hu C, Wang M, Wu C, Zhou H, Chen C, Diede S. Comparison of Duration of Response vs Conventional Response Rates and Progression-Free Survival as Efficacy End Points in Simulated Immuno-oncology Clinical Trials. JAMA Netw Open. 2021 May 3;4(5):e218175. doi: 10.1001/jamanetworkopen.2021.8175. PMID: 34047794; PMCID: PMC8164100.

[9] Huang B, Tian L, McCaw ZR, Luo X, Talukder E, Rothenberg M, Xie W, Choueiri TK, Kim DH, Wei LJ. Analysis of Response Data for Assessing Treatment Effects in Comparative Clinical Studies. Ann Intern Med. 2020 Sep 1;173(5):368-374. doi: 10.7326/M20-0104. Epub 2020 Jul 7. PMID: 32628533; PMCID: PMC7773521.

[10] Huang B, Tian L, Talukder E, Rothenberg M, Kim DH, Wei LJ. Evaluating Treatment Effect Based on Duration of Response for a Comparative Oncology Study. JAMA Oncology. 2018 6 1 ;4(6) :874 - 6 [PubMed : 29710201]

[11] Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. Nat Rev Dis Primers. 2020 Nov 26;6(1):92. doi: 10.1038/s41572-020-00224-3. Erratum in: Nat Rev Dis Primers. 2023 Jan 19;9(1):4. PMID: 33243986; PMCID: PMC7944998.

[12] Jung K, Narwal M, Min SY, Keam B, Kang H. Squamous cell carcinoma of head and neck: what internists should know. Korean J Intern Med. 2020 Sep;35(5):1031-1044. doi: 10.3904/kjim.2020.078. Epub 2020 Jul 14. PMID: 32663913; PMCID: PMC7487309.

[13] Kleinbaum, D. G., Klein, M. (2011). Survival Analysis : A Self-Learning Text, Third Edition (Statistics for Biology and Health). Springer.

[14] Luo X, Huang B, Tian L (2020). Package 'PBIR'. https ://CRAN.R-project.org/package=PBIR

[15] Marur S, Forastiere AA. Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment. Mayo Clin Proc. 2016 Mar;91(3):386-96. doi: 10.1016/j.mayocp.2015.12.017. PMID: 26944243.

[16] McCaw ZR, Tian L, Wei L. Appropriate Analysis of Duration of Response Data in Cancer Trials. JAMA Oncol. 2020;6(12):1978. doi:10.1001/jamaoncol.2020.4657

[17] Moore, D. F. (2016). Applied survival analysis using R. Switzerland: Springer.

[18] Morgan TM. Analysis of duration of response: a problem of oncology trials. Control Clin Trials. 1988 Mar;9(1):11-8. doi: 10.1016/0197-2456(88)90004-9. PMID: 3356149.

[19] Novitzke JM. The significance of clinical trials. J Vasc Interv Neurol. 2008 Jan;1(1):31. PMID: 22518214; PMCID: PMC3317309.

[20] Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol. 2016 Aug;76:175-82. doi: 10.1016/j.jclinepi.2016.02.031. Epub 2016 Mar 8. PMID: 26964707; PMCID: PMC5045274.

[21] Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med. 1998 Dec 30;17(24):2815-34. doi: 10.1002/(sici)1097-0258(19981230)17:24<2815::aid-sim110>3.0.co;2-8. Erratum in: Stat Med. 2004 Jun 15;23(11):1817. PMID: 9921604.

[22] Prinja S, Gupta N, Verma R. Censoring in clinical trials: review of survival analysis techniques. Indian J Community Med. 2010 Apr;35(2):217-21. doi: 10.4103/0970-0218.66859. PMID: 20922095; PMCID: PMC2940174.

[23] Roberts C, Torgerson D. Randomisation methods in controlled trials. BMJ. 1998 Nov 7;317(7168):1301. doi: 10.1136/bmj.317.7168.1301. PMID: 9804722; PMCID: PMC1114206.

[24] Royston, P., Parmar, M.K. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 13, 152 (2013). https://doi.org/10.1186/1471-2288-13-152

[25] Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics. 1983 Jun;39(2):499-503. PMID: 6354290.

[26] Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. Antimicrob Agents Chemother. 2004 Aug;48(8):2787-92. doi: 10.1128/AAC.48.8.2787-2792.2004. PMID: 15273082; PMCID: PMC478551.

[27] Statistical software for data science | Stata. (s.d.) https://www.stata.com/manuals13/stglossary.pdf

[28] Temkin NR. An analysis for transient states with application to tumor shrinkage. Biometrics. 1978 Dec;34(4):571-80. PMID: 571291.

[29] Tsai WY, Luo X, Crowley J The Probability of Being in Response Function and Its Applications In : Matsui S, Crowley J (eds) Frontiers of Biostatistical Methods and Applications in Clinical Oncology. 2017 Springer, Singapore.

[30] Unnebrink K, Pritsch M. Grundprinzipien klinischer Therapiestudien–was, wie und warum? [Basic principles of clinical trials–what, how, and why?]. Med Klin (Munich). 1999 Aug 15;94(8):458-64. German. doi: 10.1007/BF03044732. PMID: 10495627.

[31] Vermorken JB, Stöhlmacher-Williams J, Davidenko I, Licitra L, Winquist E, Villanueva C, Foa P, Rottey S, Skladowski K, Tahara M, Pai VR, Faivre S, Blajman CR, Forastiere AA, Stein BN, Oliner KS, Pan Z, Bach BA; SPECTRUM investigators. Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (SPECTRUM): an open-label phase 3 randomised trial. Lancet Oncol. 2013 Jul;14(8):697-710. doi: 10.1016/S1470-2045(13)70181-5. Epub 2013 Jun 6. PMID: 23746666.

[32] 14.2 Wald test | A Guide on Data Analysis. (s. d.). Home | Bookdown. https://bookdown.org/mike/data_analysis/wald-test.html

[33] World Health Organization : WHO. (2022, February, 3). Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer

[34] Clinical trials. (s. d.). World Health Organization (WHO). https://www.who.int/health-topics/clinical-trials#tab=tab_1

[35] Willems S, Schat A, van Noorden MS, Fiocco M. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. Stat Methods Med Res. 2018 Feb;27(2):323-335. doi: 10.1177/0962280216628900. Epub 2016 Mar 17. PMID: 26988930.

**UNIVERSITÉ CATHOLIQUE DE LOUVAIN**
**Faculté des sciences**
Place des Sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc