SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Medicine and Surgery

PhD program in Public Health Cycle XXXIII
Curriculum in Biostatistics and Clinical Research

Comparison of propensity score based methods for estimating marginal hazard ratios with composite unweighted and weighted endpoints: simulation study and application to hepatocellular carcinoma

Candidate: PACIFICO CLAUDIA
Registration number: 786976

Tutor: Stefania Galimberti
Co-tutor: Davide Bernasconi
Coordinator: Guido Grassi

ACADEMIC YEAR 2020/2021

Index

Introduction

A common goal of many experimental and observational clinical studies is to assess the relationship between an exposure (e.g. a treatment) and a well-defined response in a group of patients with certain characteristics. In many cases it is also of interest to establish to what extent this relationship can be interpreted as a *causal effect* of the treatment on the outcome. In the context of randomized controlled trials (RCT), this goal is made achievable by design: in particular, randomization theoretically guarantees that the possible confounders (i.e. known or unknown factors that are associated with both the treatment and the outcome) are similarly distributed among the treatment groups. This means that the contrast observed in the expected value of the outcome between groups is likely attributable to the treatment only. In the context of observational studies this issue becomes more challenging as it can be tackled only in the data analysis phase. Thus, a wide range of statistical tools are nowadays available to address this problem. Propensity score (PS) methods are increasingly being used to reduce or minimize the effect of confounding factors in observational studies of treatment effect on outcomes. Furthermore, some propensity score methods allow to estimate marginal effects rather than conditional effects provided for instance by regression models [1]. The advantage of a marginal effects, relies on its causal interpretation. The marginal effect describes the impact of treatment that could be observed in the counterfactual situation in which the all population is moved from untreated to treated. Conditional effect does not have this causal flavour, because it describes the impact of treatment in a subject with specific characteristics and not at population level.

In survival analysis the outcome is represented by time elapsed from an initial condition to the occurrence of an event of interest. A natural approach to quantify the treatment effect in survival analysis is to compare the survival or the cumulative incidence under each treatment level at some or all times t [2]. At any time t, it is also possible to calculate the incidence rate of new cases occurring in the next time-unit among those who had not yet developed the event before t. This quantity is the discrete time hazard and it may be regarded as a sort of instant velocity of event occurrence and thus it can increase or decrease over time [3]. Another frequent approach to quantify the treatment effect in survival analyses is to estimate the ratio of the hazards in the treated and the untreated, known as the hazard ratio [2]. This is frequently done using the Cox regression model, which allows estimating the hazard ratio of treatment adjusting for measured potential confounders.

Although propensity score methods have frequently been used in the analysis of time-to-event outcomes, there are few studies in literature examining the relative performance of different propensity score methods for estimating marginal hazard ratios [4] [5] [6]. In one of these studies [4], the author found that both PS matching and Inverse Probability Weighting (IPW) allow for the estimation of marginal hazard ratios with minimal bias, while stratification on the PS and covariate adjustment using the PS result in biased estimation of marginal hazard ratios. Considering these results, the author suggested to use PS matching and IPW when the interest is to estimate the relative effect of treatment on time-to-event outcomes. Another study [5] demonstrates that a failure to account for the sampling variability can bring to incorrect statistical inference when PS weighting analysis is performed, while [6] shows that an estimator based on bootstrap resampling provides a good approximation of standard errors and thus an adequate coverage of the confidence interval for marginal hazard ratios [6]. The results presented in the above-mentioned studies all rely on a broad series of Monte Carlo simulations. However, within this framework, some of the settings chosen (e.g. limited number of confounders and all with a Gaussian distribution, absence of unmeasured confounders and correct PS model specification, extremely high sample size) appear unrealistic and not representative of the practical issues often arising in practice when analysing data from an observational study. Another issue that commonly arises when dealing with survival outcomes consist in analysing composite endpoint, that combine several specific event. The analysis of composite endpoint is straightforward because standard statistical method can be used; however, sometimes it is of interest to assign a different clinical relevance to each cause-specific event and ad hoc measure is been proposed to this aim. For example, one idea is to focus on a weighted composite survival endpoint [7].

The methodological aim of this thesis is the comparison of the performance of different PS based methods, through simulation studies, in estimating the marginal effect of treatment on standard (unweighted) or weighted composite survival endpoints. The study of the causal effect of treatment on weighted endpoints is a completely innovative aspect, as currently there are no references in the literature about this topic.

The motivating clinical study of this thesis is part of the HERCOLES project (Hepatocarcinoma Recurrence on the Liver Study), an Italian retrospective study on hepatocellular carcinoma [8]. The primary aim of this study is to compare the prognosis of patients with primary hepatocarcinoma undergoing different surgical techniques (anatomic resection vs wedge resection). The main endpoint of interest is the disease free survival defined as time from treatment to the first of the

following specific events: local hepatic recurrence (i.e. recurrence on the surgical cut of the liver), non-local hepatic recurrence and death without recurrence. We analysed the causal effect of treatment of the composite endpoint in term of marginal hazard ratio by combining several PS-based methods with Cox regression. Furthermore, it is of interest to consider the different clinical relevance of the cause-specific events. In particular, in the clinical practice, death is considered the worst event, but also more relevance is given to local recurrence compared to the non-local one. To account for the different relevance assigned to each endpoint we considered the method of Ozga and Rauch who recently proposed a new non-parametric weighted effect measure and estimator for composite endpoints called the 'weighted all-cause hazard ratio' [9]. We used this method for the analysis of this study, in combination with IPW and PS-matching, to estimate a causal effect of treatment on the weighted composite endpoint.

Regarding the structure of this thesis, in Chapter 1 the statistical methods relevant for the purposes of our study are introduced. Specifically, a description of the basic concepts of survival analysis and a definition of the standard and weighted all-cause hazard-ratio is given. Furthermore, an introduction to causal inference and a description of the different PS based methods are provided. In Chapter 2 and 3 the simulation studies on standard and weighted composite endpoints are described, respectively. In light of the results obtained, in Chapter 4 the clinical application of these methods on the HERCOLES data is reported. In the discussion (Chapter 5) all the findings are summarized and placed in the context of the existing literature.

1. Methods

1.1. Standard survival analysis

Survival analysis is a collection of statistical methods for data analysis for which the outcome variable of interest is time elapsed from a starting point (e.g. diagnosis of disease, administration of treatment, beginning of exposure) until an event occurs (e.g., death, disease incidence, relapse from remission, recovery). This time is often called "survival time" even when the event of interest is not death. Time can be measured in years, months, weeks, or days depending on the type of event of interest.

A typical problem in survival analysis is the presence of *censoring*: for some individuals the time to occurrence of the event of interest is only partially available.

There are three types of censoring:

1. Right censoring: true survival time is equal to or greater than observed survival time
2. Left censoring: true survival time is less than or equal to the observed survival time
3. Interval censoring: true survival time is within a known time interval

The most common is right censoring and the reasons why it may occur are:

1. The study ends before a subject experiences the event (i.e. administrative censoring)
2. A subject is lost to follow-up during the study
3. A subject withdraws the treatment (drop out or treatment abandon), this may occur especially in RCT

Theoretical functions in survival analysis

Let T denote the survival time. Beside the density function, two other functions are commonly used in survival analysis to describe the distribution of T : the survival function, denoted by $S(t)$, and the hazard function, denoted by $h(t)$. The survival function is defined as the probability that a subject survives longer than t:

$$S(t) = P(T > t) \quad (1)$$

Theoretically, as t ranges from 0 up to infinity, the survival function can be graphed as a smooth curve:

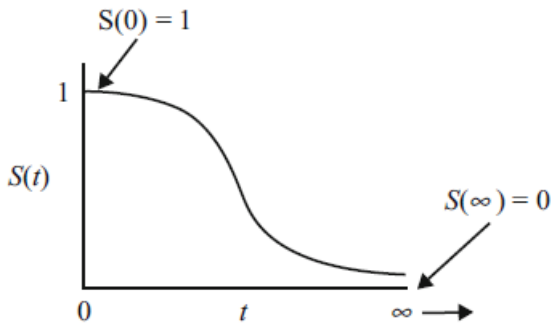


Figure 1. Survival function smooth curve

Moreover, $S(t)$ is a non-increasing function of time *t* with the following properties:

$$S(t) = \begin{cases} 1 & for\ t = 0 \\ 0 & for\ t = \infty \end{cases} \quad (2)$$

The complement to 1 of the survival function is the cumulative incidence function, which represents the probability of event occurrence at time *t* or before:

$$F(t) = 1 - S(t) = P(T \leq t) \quad (3)$$

The hazard function, denoted by $h(t)$, is given by the following formula:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (4)$$

The hazard function $h(t)$ gives the instantaneous event rate which describes the risk at time *t* that the event of interest occurs in the next time unit, given that the individual has survived up to time *t* [10]. The cumulative hazard function is the integral of the hazard function between integration limits of 0 and t:

$$H(t) = \int_0^t h(u)du \quad t > 0 \quad (5)$$

The survival function can also be expressed in terms of the hazard function by the negative exponent of the cumulative hazard function:

$$S(t) = e^{-H(t)} \quad t > 0 \quad (6)$$

From this one-to-one relationships, it is possible to obtain the relationship between $h(t)$ and $S(t)$:

$$h(t) = -[\frac{dS(t)/dt}{S(t)}] \quad (7)$$

Considering a parametric survival model, time is assumed to follow some distribution whose probability density function $f(t)$ can be expressed in terms of unknown parameters. Once $f(t)$ is specified, the corresponding survival and hazard functions can be determined. The $S(t)$ can be obtained as follows:

$$S(t) = \mathrm{P}(\mathrm{T} > \mathrm{t}) = \int_t^{\infty} f(u)du \quad (8)$$

The hazard can then be found from (7).

Estimators in survival analysis

The Kaplan-Meier (KM) estimator [2], also known as the product limit estimator, can be used to estimate the survival function from survival data in presence of censored data, assuming independent censoring (i.e. censoring time is independent from the true survival time). This estimator is the non-parametric maximum likelihood estimator of $S(t)$ and can be expressed as follows:

$$\widehat{S}(t) = \prod_{ti<t} \frac{n_i - d_i}{n_i} \quad (9)$$

At each event time t_i there are n_i subjects "at risk" and d_i number of deaths/failures. Censored individuals before time t_i are not anymore in the risk set n_i.

In large samples, $\hat{S}(t)$ is approximately normally distributed with mean $S(t)$ and a variance which may be estimated by Greenwood's formula:

$$Var\big(\hat{S}(t)\big) = \hat{S}(t)^2 \sum_{ti<t} \frac{d_i}{n_i(n_i - d_i)} \quad (10)$$

An important advantage of the KM curve is that this method is robust for right censoring. When no truncation or censoring occurs, the KM curve is the complement of the empirical distribution function. Nevertheless, there are a lot of different situations where the KM cannot be used. For example, when more than one type of event is considered (i.e. competing risks analysis) the incidence function of each specific event (*crude incidence*) is the quantity of interest and a valid estimator for this function must be used (see next paragraph).

The Nelson-Aalen estimator [11] is a non-parametric estimator of the cumulative hazard, assuming independent censoring. It is expressed by the following formula:

$$\hat{A}(t) = \sum_{t_j \le t} \frac{d_j}{n_j} \quad (11)$$

The Nelson–Aalen estimator is an increasing right continuous step function with increments $\frac{d_j}{n_j}$ at the observed failure times. The variance of the Nelson–Aalen estimator may be estimated by:

$$\widehat{\sigma^2}(t) = \sum_{t_j \le t} \frac{(r_j - d_j)d_j}{(r_j - 1)n_j^2} \quad (12)$$

The Cox Proportional Hazards (PH) model [2], most commonly known as the Cox model, is a semi-parametric regression method that can be used to analyse the impact of multiple covariates on a survival outcome. The model formula for the hazard is the following:

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i} \quad (13)$$

This model gives an expression of the hazard at time *t* for an individual with a given collection of p explanatory variables (X) and requires the assumption of independent censoring conditional on covariates [1] [10].

In the Cox model formula, the hazard at time *t* is the product of two quantities:

i. $h_0(t)$: the baseline hazard function, that is a function of *t* that does not involve the X's

ii. $e^{\sum_{i=1}^{p} \beta_i X_i}$: a function of X that does not involve *t*

These properties lead to the proportional hazards (PH) assumption that characterizes the Cox model. One of the main reasons for the popularity of the Cox model is that, even though the baseline hazard is not specified, reasonably good estimates of regression coefficients and adjusted survival curves can be obtained in many situations.

The regression coefficient of the Cox model is the Hazard Ratio (HR): the hazard for one individual divided by the hazard for a different individual. The two individuals compared can be differentiated by their X's values. The HR can be written as the estimate of $\hat{h}(t, X^*)$ divided by the estimate of $h(t, X)$, where X^* denotes the set of predictors for one individual, and X denotes the set of predictors for the other individual.

$$\widehat{HR} = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)} \quad (14)$$

An expression for the HR formula in terms of the regression coefficients is obtained by substituting the Cox model formula into the numerator and denominator of the HR expression.

$$\widehat{HR} = \frac{\hat{h}_0(t)e^{\sum_{i=1}^{p}\widehat{\beta_i}X_i^*}}{\hat{h}_0(t)e^{\sum_{i=1}^{p}\widehat{\beta_i}X_i}} = e^{\sum_{i=1}^{p}\widehat{\beta_i}(X_i^* - X_i)} \quad (15)$$

Hence, HR is calculated without having to estimate the baseline hazard function and then it is not time-dependent. When there is only one binary covariate $X_1 \in \{0;1\}$ (e.g. an indicator of treatment or exposure) the estimated HR reduces to e^{β_1}.

Estimators in survival analysis with competing risk

In case of competing risks, several types of event may originate the failure time T and are thought as competing causes. In this context, the quantities of interest are the cumulative incidence of any event and also the incidence of each specific type of event and its contribution to the overall incidence. The incidence of each type of event (also called "crude incidence") can be estimated using the Aalen-Johansen estimator [12]. It is a standard non parametric method to estimate the cumulative incidence, generalizing the Kaplan-Meier estimator to multiple event types under the assumption of independent censoring. This estimator is the sum of unconditional probabilities of failure due to the event of interest in time, obtained by multiplying the probability of having survived any event by the cause-specific hazard of the event of interest:

$$\hat{F}_1(t) = \prod_{ti \le t} \hat{h}_1(t_i) \cdot S(t_i) \quad (16)$$

Where:

$$\hat{h}_1(u) = \frac{d_{1i}}{n_i} \quad (17)$$

Where d_{1i} is the number of events of type 1 at t_i and n_i the total of individuals at risk at t_i. Of note, this estimator is not equivalent to apply the Kaplan-Meier estimator after censoring the observations at the times of all competing events. In fact, this procedure would lead to overestimate the crude incidence.

1.2. Survival analysis for composite endpoint

1.2.1. The standard all-cause hazard ratio

Composite Endpoints (CE) combine several events within a single variable, which means that the time to occurrence of the first among different events is considered. The rationale for the use of composite time-to-event endpoints is to increase the number of expected events and thereby the power by combining several event types of clinical interest [13]. A common example of CE is the disease free survival, which is the time until relapse or death without prior relapse, whatever occurs first. The methods used to analyse a CE and to evaluate how they are affected by a certain treatment or exposure are the same as those described for the standard survival analysis.

Thus in the univariate case of a single dichotomous variable X_i which is equal to 1 when the subject *i* belongs to the intervention (I) group and 0 when it belongs to the control (C), the hazard of the composite endpoint can be modelled as:

$$\lambda_{CE,i}(t) = \lambda_{CE,0}(t)e^{\beta_{CE}X_i} \qquad (18)$$

This function is also called "all-cause hazard" and can be reworded as the sum of the cause-specific hazards for the *k* single endpoints (EP_j):

$$\lambda_{CE}(t) = \sum_{j=1}^{k} \lambda_{EP_j}(t) \qquad (19)$$

As a consequence, the standard "all-cause hazard ratio", i.e. the hazard ratio of the CE, is given as

$$\theta_{CE} = e^{\beta_{CE}} = \frac{\lambda_{CE}^{I}(t)}{\lambda_{CE}^{C}(t)} \qquad (20)$$

Of note, the formula used in (20) indicates that the proportional hazards assumption is assumed, meaning that θ_{CE} is constant in time. Thus, the standard Cox model can be adopted to estimate θ_{CE}.

1.2.2. The weighted all-cause hazard ratio

Using a composite time-to-event endpoint, the effect of the individual components might differ, in magnitude or even in direction, which leads to interpretation difficulties. Moreover, the individual event types often are of different clinical relevance which further complicates interpretation.

Starting from this statements, Rauch et al. [7] introduced the idea of the weighted all-cause hazard to replace the standard one (19) with a weighted sum of the cause-specific hazards using fixed weights. Hence the weighted all-cause hazard is given as

$$\lambda_{CE}^{w}(t) = \sum_{j=1}^{k} W_{EP_j} \lambda_{EP_j}(t) \quad (21)$$

The non-negative weights $w_{EP_j} \geq 0, j = 1, \ldots, k$, reflect the clinical relevance of the components $EP_j,\ j = 1, \ldots, k$. If all the weights were equally set to 1, then the weighted all-cause hazard would correspond to the standard all-cause hazard.

The "weighted all-cause hazard ratio" as proposed by Rauch et al. [7] is then given as

$$\theta_{CE}^{w}(t) = \frac{\lambda_{CE}^{I,w}(t)}{\lambda_{CE}^{C,w}(t)} = \frac{\sum_{j=1}^{k} w_{EP_j} \lambda_{EP_j}^{I}(t)}{\sum_{j=1}^{k} w_{EP_j} \lambda_{EP_j}^{C}(t)} \quad (22)$$

To obtain an estimate of $\theta_{CE}^{w}(t)$, Rauch et al. [7] proposed to estimate each cause-specific hazard via a parametric survival model (e.g. Weibull model) and to plug in the estimate of each λ_{EP_j} in (22). Following this approach, the weighted all-cause hazard ratio is given by

$$\hat{\theta}_{CE}^{w}(t) = \frac{\sum_{j=1}^{k} w_{EP_j} \hat{\lambda}_{EP_j}^{I}(t)}{\sum_{j=1}^{k} w_{EP_j} \hat{\lambda}_{EP_j}^{C}(t)} \quad (23)$$

Of note, a variance estimator for (23) cannot easily be measured and so an asymptotic distribution of the parametric estimator given in (23) is not available.

Moreover, since the shape of the survival distribution is usually not known in advance, its pre-specification, due to the choice of a parametric survival model, must be seen as a strong restriction. Thus, there is the general interest in deriving a more flexible non-parametric estimator.

To derive a non-parametric estimator, Ozga et al. [9] propose to replace the cause-specific hazards by the cumulative cause-specific hazards:

$$\theta_{CE}^{w}(t) = \frac{\sum_{j=1}^{k} w_{EP_j} \Lambda_{EP_j}^{I}(t)}{\sum_{j=1}^{k} w_{EP_j} \Lambda_{EP_j}^{C}(t)} \quad (24)$$

where $\Lambda_{EP_j}(t),\ j = 1, \ldots, k,$ refer to the corresponding cause-specific cumulative hazards over the period [0, t]. Using the non-parametric Nelson-Aalen estimators, it is possible to derive a non-parametric estimator for the weighted all-cause hazard ratio:

$$\hat{\Lambda}^I_{EP_j}(t) := \sum_{t_l \le t} \frac{d^I_{EP_j,l}}{n^I_l}, \quad \hat{\Lambda}^C_{EP_j}(t) := \sum_{t_l \le t} \frac{d^C_{EP_j,l}}{n^C_l} \qquad (25)$$

By this, a non-parametric estimator for the weighted all-cause hazard ratio is given by

$$\tilde{\theta}^w_{CE}(t) = \frac{\sum_{j=1}^k w_{EP_j} \widehat{\Lambda}^I_{EP_j}(t)}{\sum_{j=1}^k w_{EP_j} \widehat{\Lambda}^C_{EP_j}(t)} \qquad (26)$$

However, since the quantity of interest is the instantaneous cause-specific hazard, to ensure that the ratios between cumulative hazards and instantaneous hazards are the same, the authors assume that the cause-specific hazards are proportional:

$$\lambda_{CE,i}(t) = \lambda_0(t) \sum_{j=1}^k e^{\beta_{EP_j} X_i} \qquad (27)$$

The proportional hazard assumption is verified if the baseline hazards for the *k* components (i.e. events) are equivalent within each group, meaning that the weighted all-cause hazard ratio is no longer time-dependent as:

$$\theta^w_{CE} = \frac{\sum_{j=1}^k w_{EP_j} \lambda^I_{EP_j}(t)}{\sum_{j=1}^k w_{EP_j} \lambda^C_{EP_j}(t)} = \frac{\sum_{j=1}^k w_{EP_j} \lambda_0(t) e^{\beta_{EP_j} * 1}}{\sum_{j=1}^k w_{EP_j} \lambda_0(t) e^{\beta_{EP_j} * 0}} \qquad (28)$$

Of note, in formula (28) the baseline hazard λ_0 cancels out because it is assumed the same for all components. As mentioned, the correctness of the non-parametric estimator is based on the assumption of equal cause-specific baseline hazards. Theoretically, in case the baseline hazards differ, $\tilde{\theta}^w_{CE}(t)$ can be calculated but represents a biased estimator for $\theta^w_{CE}(t)$. However, Ozga et al. [9] show that in practice the estimator is remarkably robust with respect to this assumption and provides fairly unbiased estimates even in the case in which the baseline hazards are not equal.

Furthermore, another restriction is due to the fact that the only possibility to calculate standard errors, and thus confidence intervals, for the weighted hazard ratio is by means of resampling methods (e.g. bootstrap).

Guidance for the choice of weights

The use of the weighted all-cause hazard ratio requires to fix the weights in the planning stage of the study. Introducing the component weights in (21), the event time distribution, that is the corresponding survival function, is implicitly modified. If the chosen weight is unequal to 1, the shape of the survival distribution changes: for a weight greater than 1, the number of events artificially increases and as a consequence, the survival function decreases sooner; for a weight smaller than 1, the number of events decreases and so the survival distribution becomes more flat. Whereas the all-cause hazard ratio can be heavily masked by a large cause-specific hazard of a less relevant component, a more relevant component with a lower number of events can only have a meaningful influence on the composite effect measure, when it is up-weighted (or if the less relevant component is down-weighted accordingly). On the contrary, if a large cause-specific hazard is down-weighted this can result in a power loss. Therefore, weighting can improve interpretation but the effect on power can be positive or negative, depending on the data at hand.

To help researchers with this task, Ozga et al. [9] provide some guidelines on how to choose appropriate weights. At first, the authors suggest to identify the clinically most relevant event type (e.g. death), assigning a weight of 1 to it. The weights for the remaining events are chosen based on the relative clinical relevance with respect to event with weight of 1. For example, one can think about the number of events of interest that could be considered as equally harmful as one event of the clinically most relevant endpoint. The latter information is difficult to obtain and it often needs the contribution of both clinical knowledge and data support to be elicited. Another recommendation consists in considering different clinically meaningful weighting schemes in order to evaluate the results in different scenarios.

1.3. Propensity score based methods

Over the past decades, several studies have highlighted the need to define more appropriately concepts as association and causality relationships: the association is a relationship without a necessary specific direction (undirected), while causality is characterized by a specific direction.
During these years, many statisticians have spent time analysing this concept in order to develop models to perform causal inference [3].
Considering for example a binary exposure A and an outcome also binary Y. From a probabilistic point of view, the association between A and Y can be defined by:

$$P[Y = 1|A = 1] \neq P[Y = 1|A = 0] \quad (29)$$

Now consider the two random variables $Y_{a=1}$ e $Y_{a=0}$ which represent the outcomes that I would have observed if I could have submitted the entire population to both treatments. These variables are called counterfactual outcomes since only one of the two is observed for each subject (the factual outcome).
The exposure has a causal effect on the outcome (in the binary example) if for each subject:

$$Y_{a=1} \neq Y_{a=0} \quad (30)$$

Since the non-factual outcome is not observable at the individual level, it is necessary to define the causal effect at the population level. The exposure A has an average causal effect on the outcome Y in the entire population if:

$$E[Y_a] \neq E\left[Y_{a^i}\right] \quad (31)$$

for each pair a, a^i, with $a \neq a^i$.
Hence, in the binary example this condition becomes:

$$P[Y_{a=1} = 1] \neq P[Y_{a=0} = 1] \quad (32)$$

The effect measures used to measure any causal relationship compare what would happen in a population under two possible but distinct scenarios, of which at most one can occur. For this reason, these measures cannot be measured directly from the data.
The only condition in which it is possible to make consistent estimate of causal measures is the exchangeability:

$$P[Y_a = 1|A = 1] = P[Y_a = 1|A = 0] = P[Y_a = 1] \quad (33)$$

valid only for randomized experiments, where the assignment of treatments occurs randomly.

On the other hand, the observational studies may have a different distribution of some prognostic factors between the two treatment groups that leads to the lack of exchangeability between exposed and unexposed. However, under suitable assumptions, also in observational studies it is possible to estimate causality measures.

The first assumption is the *conditional exchangeability* described as follows:

$$P[Y_a = 1|A = 1, L = l] = P[Y_a = 1|A = 0, L = l] = P[Y_a = 1|L = l] \quad (34)$$

where A represents the exposure, Y the outcome, Y_a the counterfactual outcome and L represents the predictive factor.

In epidemiological terms, the predictive factor L can be considered as a confounding, that is, a variable that is associated to both the exposure A and outcome Y but that does not appear in the causal path of the exposure-outcome relationship.

The other two assumptions that must be satisfied are the *consistency* and the *condition of positivity*. The *consistency* is defined as $Y_a = Y$ if A=a is the treatment actually received by the subject; while the *positivity* condition implies:

$$P[A = a|L = l] > 0 \quad if\ P[L = l] \neq 0 \quad (35)$$

this assumption means that it must be ensured (from the study design) that there is a probability greater than zero of being assigned to each of the treatment levels.

Therefore, the ability to identify a causal effect from an observational study depends on whether the confounding L is measured. Many statistical methods used to remove or reduce the effect of the different distribution of the confounders between the two treatment groups are based on the concept of PS.

The PS is defined as the probability that the *i*-th subject is assigned to a treatment conditional on confounders measured at baseline:

$$PS = Pr(A_i = 1|X_{1i}, X_{2i}, \dots, X_{ki}) \quad (36)$$

In absence of randomization, balancing of the PS between treatment groups guarantees the conditional exchangeability which is essential to quantify a causal measure.

The estimation of the PS is typically performed by logistic regression, where the binary outcome is represented by the treatment indicator and the potential measured confounders are included in the model as covariates.

The effect of treatment evaluated using a multiple regression approach (adjusting for potential confounders) is regarded as a *conditional effect*: the effect at subject level of moving a single subject from untreated to treated. The aim of the PS based methods is to estimate the *marginal effect*: the effect at population level, of moving an entire population from untreated to treated. In a causal notation, this effect is called the Average Treatment Effect (ATE). A slightly different situation occurs when the interest is in studying the effect of moving only the actually treated population from untreated to treated. In this case, the estimated effect is called the Average Treatment Effect on the Treated (ATT). Some PS based methods cannot produce unbiased estimates for ATE nor ATT, because they focus on a conditional effect. However, all these methods are used in applied literature to reduce or minimize confounding aiming to obtain an effect that could be interpreted as causal. In this work, the hazard ratio is the measure used to quantify the effect of treatment. An additional complication of dealing with such quantity in a causal perspective is that, differently from other effect measures such as Risk Difference or Relative Risk, the HR is not collapsible: the conditional and marginal treatment effects do not coincide even in the absence of confounding [4][5]. In the next paragraph, some well-known PS based methods are described, including both methods for marginal and conditional effect.

1.3.1. Matching on propensity score

It has been shown that matching on PS theoretically gives unbiased estimates of the ATT [20], which in principle can be different from ATE. However, in this work, we analysed the performance of this method in order to estimate a marginal HR which represents the ATE on unweighted and weighted composite survival endpoints. This is motivated by the fact that the ATE is the true quantity of interest in many applied studies and that in practice ATE and ATT could be very similar.

The aim of the PS matching is to form matched sets of treated and untreated subjects who have a similar value of the PS. In literature, there are different algorithms for forming pairs of treated and untreated subjects matched on the propensity score.

The most common algorithms are optimal matching, nearest neighbour matching with or without replacement and greedy nearest neighbour matching with or without replacement within specified caliper widths. In the matching without replacement we matched each untreated subject to at most one treated subject, while in the matching with replacement the same untreated subject can be matched to multiple treated subjects. *Optimal matching* forms matched pairs so as to minimize the

average within-pair difference in propensity scores. In contrast, the *nearest neighbour matching* is simply performed by matching each treated subject to the untreated subject whose propensity score is closest to that of the treated subject. A more refined procedure is called *greedy nearest neighbour caliper matching* and consists in matching treated and untreated subjects only if the absolute difference in their propensity scores is within a pre-specified maximal distance (the caliper distance). When using caliper matching, we match subjects on the logit of the propensity score using a caliper width that is defined as a proportion of the standard deviation of the logit of the propensity score [14][15][16]. A caliper of width equal to 0.2 standard deviations of the logit of the propensity score, has been found to perform well in a wide variety of settings [17].

Once a PS-matched sample has been formed, it is possible to estimate marginal survival functions (using the Kaplan–Meier estimator) and the marginal hazard ratio between treated and untreated for the composite unweighted endpoint. This is simply done by fitting a Cox model to the PS-matched sample including the treatment indicator as the only covariate and taking the exponential of the corresponding coefficient, as in formula (15). Concerning the all-cause weighted endpoint, one can estimate the marginal hazard ratio simply applying the estimator of Ozga et al. defined in (26) to the PS-matched sample.

1.3.2. Stratification on the propensity score

Stratification on the PS is not appropriate to estimate marginal effect, because focuses on a conditional measure similarly to a standard regression approach.

The method consists on splitting out the entire sample into mutually exclusive subclasses based on the propensity score. In literature, the most popular approach is to define the subclasses using specified quantiles of the PS (typically quartiles or quintiles). Besides the HR, this method can be used to estimate adjusted survival curves for each of the two treatment groups through the Kaplan–Meier estimator. When estimating the conditional effect, each stratum is weighted proportionally to the number of treated subjects who lay within that stratum. Essentially, one is pooling stratum-specific survival curves to obtain a population-average survival curve. In this work, stratification on PS is used in order to estimate an adjusted HR only for the unweighted composite survival endpoint. This is done by fitting a Cox model on the original sample and including as covariates the treatment indicator and also the PS-stratum indicator. Again, the hazard ratio is obtained by taking the exponential of the treatment coefficient.

1.3.3. Propensity score as covariate in a model

Including the PS in a regression model is another method theoretically focusing on conditional rather than marginal effects. A regression model for the study outcome can directly include the treatment indicator and the PS as explanatory variables. Flexible methods for the transformation of the PS variable should be considered, such as cubic splines or fractional polynomials, as the association between the score and the outcome may not be linear [18]. Again, the adjusted HR is obtained by calculating the treatment HR as the exponential of the treatment coefficient on the whole set of data.

Another, less parsimonious, variation is to include, on the top of the treatment variable, both PS and important confounding variables as covariates in the regression model.

In this work, PS as a covariate is used in order to estimate the adjusted HR only for the unweighted composite survival endpoint.

1.3.4. Inverse probability weighting

The aim of the inverse probability weighting (IPW) method is to create a hypothetical population in which every individual appears as a treated and as an untreated individual.

Hence it is possible to quantify the ATE: the treatment effect that would be observed if the entire population could be submitted to both treatments.

The size of the pseudo-population can be calculated by multiplying the observed numbers by the following weights:

$$W = \frac{1}{P(A|L)} \quad (37)$$

A simple numerical example is the following [19].

Considering a population in which twenty subjects are followed over time to evaluate the effect of smoke on cardiovascular event. Of these, twelve have a high stress state (L=1) and of these nine smoke (A=1). Between these are observed six cases of cardiovascular event (Y=1) while among the subjects with high level of non-smoking stress two cases are observed. Between the eight subjects whose state of stress is not high, four are smokers and, of these, one experienced a cardiovascular event. Among the remaining four subjects (non-smokers) one case of cardiovascular event is observed. The data are summarized in figure 2.

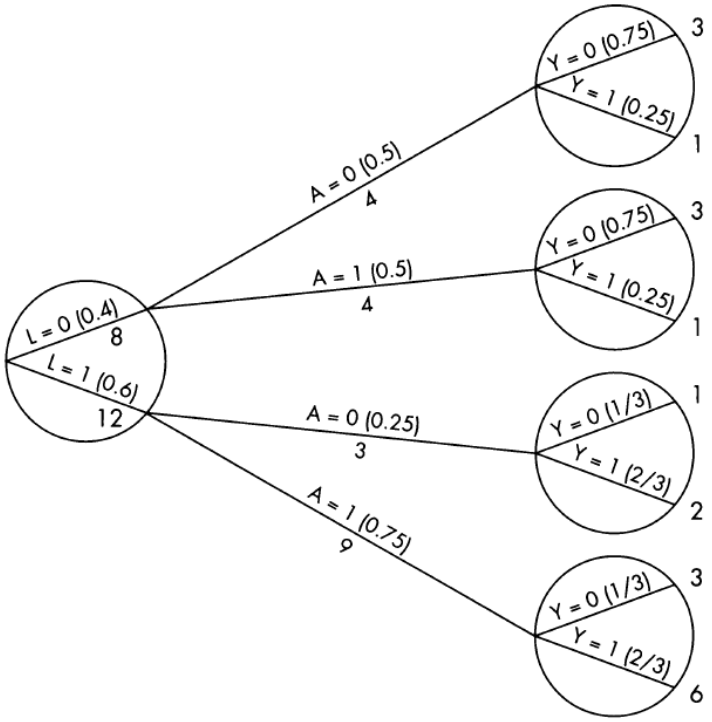


Figure 2. A population with confounders L, exposure A and outcome Y

Figure 3 represents the pseudo-population obtained if all subjects of the original population had been non-smokers (i.e. untreated).

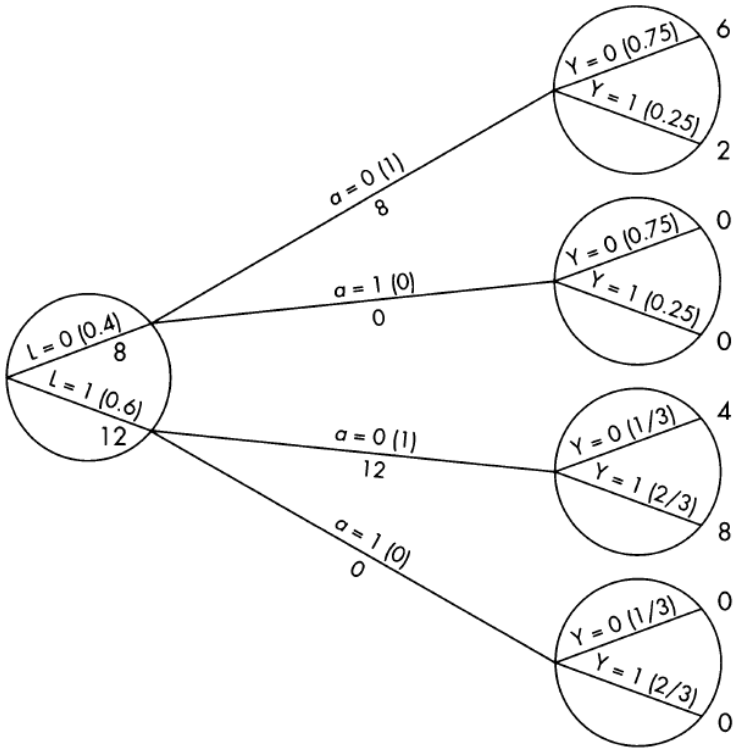


Figure 3. The population had everybody remained unexposed

The number of branches corresponding to Y = 0 and Y = 1 reflect the original proportions. Similarly, the tree corresponding to the pseudo population obtained if all subjects had been smokers (i.e. treated). This population is represented in figure 4.

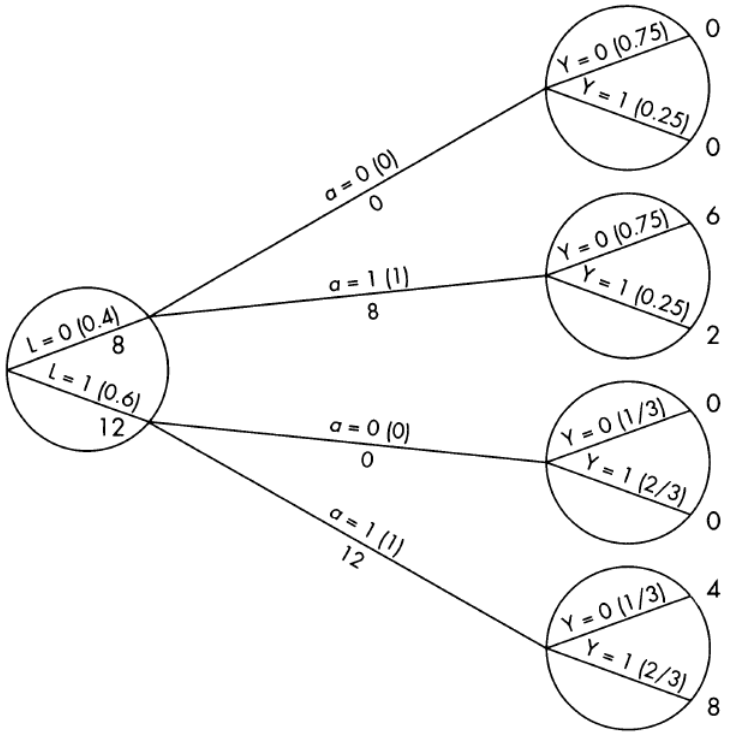


Figure 4. The population had everybody remained exposed

By merging the previous two pseudo populations, the overall pseudo population is obtained. This pseudo-population can be obtained by applying the weights defined by (37). In figure 5 represents the overall pseudo population:

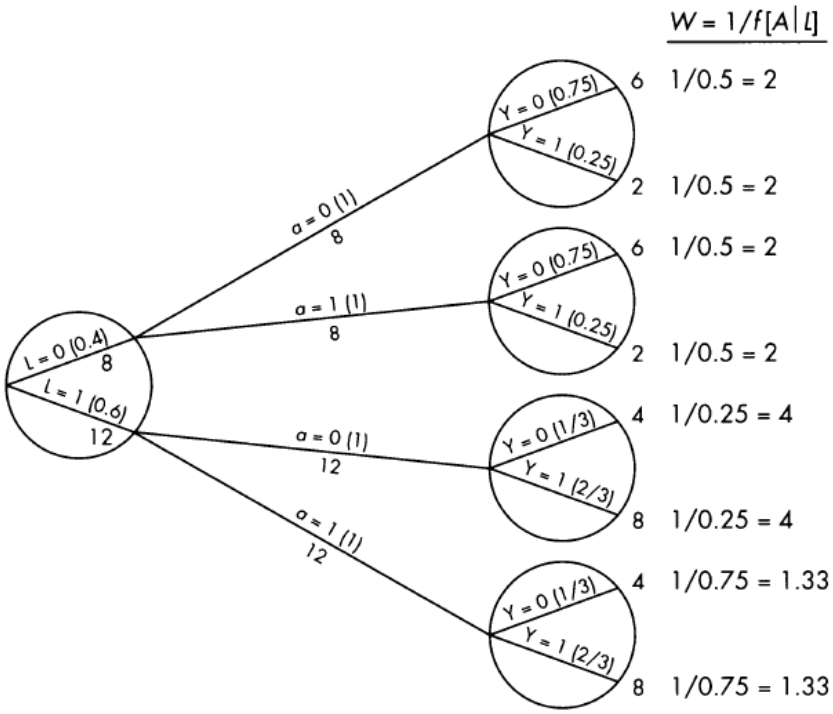


Figure 5. The overall pseudo population

For example, the first set of subjects having A = 0 and L = 0 corresponds to a weight equal to:

$$W = \frac{1}{P(A=0|L=0)} = \frac{1}{P(A=0,L=0)/P(L=0)} = \frac{P(L=0)}{P(A=0,L=0)} = \frac{8/20}{4/20} = 2$$

and consequently the corresponding pseudo population size will be equal to 2*4 (weight* original population size). In this way the size of the pseudo population increases (doubles).

If instead of considering the weights W, the following stabilized weights are implemented:

$$SW = \frac{P(A)}{P(A|L)} \quad (38)$$

the size of the pseudo-population remains the same as the starting one since the numerator represents a distribution whose integral is 1. In complex situations it is preferable to use stabilized weights than the non-stabilized one, since this makes the estimator more efficient.

The calculation of the weights is done in practice using the logistic regression estimates of the PS to calculate the denominator. In the case of stabilized weights, the numerator can be obtained using an empty (i.e. without covariates) logistic regression model where treatment status represents the

outcome. In this work, the IPW is used in order to estimate the marginal HR of treatment on unweighted and weighted composite survival endpoints from the pseudo population.

Once a pseudo-population has been formed, it is possible to estimate marginal survival functions (using the weighted Kaplan–Meier estimator) and the marginal hazard ratio between treated and untreated for the composite unweighted endpoint. This is simply done by fitting a Cox model to the pseudo-population (i.e. weighting each observation using inverse probability weights) including the treatment indicator as the only covariate and taking the exponential of the corresponding coefficient, as in formula (15). Concerning the all-cause weighted endpoint, one can estimate the marginal hazard ratio simply applying the estimator of Ozga et al. defined in (38) to the pseudo-population (again, weighting each observation using inverse probability weights).

2. Simulations on unweighted hazard ratio

2.1. Simulation protocol

Data generation (i.e. the number of covariates, their distribution, their association with the treatment and the outcome, the distribution of treatments and the sample size) was inspired by data observed in the HERCOLES study (see Chapter 4) in order to mimic a situation which can occur in practice. The twelve baseline covariates (X_1-X_{12}) are simulated from different distributions (Normal, Binomial and Poisson) whose parameters are inspired by the distribution of the confounders identified for the HERCOLES study:

- $x_1 \sim Bin(1; 0.7)$
- $x_2 \sim Poisson(1.3)$
- $x_3 \sim N(75; 10)$
- $x_4 \sim Bin(1; 0.1)$
- $x_5 \sim Bin(1; 0.51)$
- $x_6 \sim N(1.2; 0.24)$
- $x_7 \sim Poisson(180)$
- $x_8 \sim N(4.6; 2.9)$
- $x_9 \sim Bin(1; 0.66)$
- $x_{10} \sim Bin(1; 0.73)$
- $x_{11} \sim Bin(1; 0.02)$
- $x_{12} \sim Bin(1; 0.48)$

For the *i-th* subject, the probability of being assigned to one of two treatments is determined from the following logistic model:

$$\boldsymbol{l}ogit(p_i) = \alpha_{0,treat} + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_{12} x_{12} \quad (39)$$

The intercept of the model ($\alpha_{0,treat}$) is set equal to 3.2 in order to obtain the desired proportion of treated subjects equal to 38%.

The regression coefficients $\alpha_1 - \alpha_{12}$ have been set according to the association with the treatment observed on the HERCOLES dataset:

- $\alpha_1 = 0.53$
- $\alpha_2 = 0.38$

- $\alpha_3 = -0.01$
- $\alpha_4 = 0.49$
- $\alpha_5 = -0.08$
- $\alpha_6 = -1$
- $\alpha_7 = -0.01$
- $\alpha_8 = -0.23$
- $\alpha_9 = -0.35$
- $\alpha_{10} = 0.02$
- $\alpha_{11} = 0.77$
- $\alpha_{12} = 0.59$

The treatment status is generated from a Bernoulli distribution with subject-specific parameter p_i and the outcome is generated using a data-generating process for time to-event outcomes described by Bender et al [20]. A linear predictor is defined for the *i-th* subject as:

$$LP = \beta_{treat}Z + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{12} x_{12} \quad (40)$$

A random number is generated from a standard uniform distribution $u_1 \sim U(0,1)$ and the censorship are generated from the following uniform distribution $u_2 \sim U(360,500)$.

The regression coefficients $\beta_1 - \beta_{12}$ have been set according to the association with the treatment observed on the HERCOLES dataset:

- $\beta_1 = 0.19$
- $\beta_2 = 0.22$
- $\beta_3 = -0.01$
- $\beta_4 = 0.16$
- $\beta_5 = -0.13$
- $\beta_6 = -1.78$
- $\beta_7 = -0.01$
- $\beta_8 = -0.02$
- $\beta_9 = 0.72$
- $\beta_{10} = 1.49$
- $\beta_{11} = 0.65$
- $\beta_{12} = 0.39$

The survival time for each subject is generated from a Weibull distribution as follows:

$$T = -\log(u)/(\gamma e^{LP})^{1/\eta}\ ,\ with\ \gamma = 0.01\ and\ \eta = 0.8 \quad (41)$$

The corresponding hazard function is:

$$\lambda(t) = \lambda_0(t)e^{LP} \qquad with\ \lambda_0(t) = \eta\gamma^{\eta}t^{\eta-1} \quad (42)$$

The parameters β_i, λ and η are set in order to obtain a survival times distribution similar to the HERCOLES dataset. This data-generating process results in a conditional treatment effect, with a conditional hazard ratio of $e^{\beta_{treat}}$. However, as the aim is to generate data from a specified marginal hazard ratio, an iterative process proposed by Austin et al. [4] was used to obtain the value of β_{treat} that induced the desired marginal hazard ratio. This process consists on the following steps:

1. Fixing the desired marginal HR
2. Calculation of the marginal HR induced by an hypothetical β_{treat} (i.e. the logarithm of the conditional HR)
3. Comparison between the calculated marginal HR and the desired marginal HR:
   a. If marginal HR = marginal HR desired, the β_{treat} value is found
   b. If marginal HR < marginal HR desired , β_{treat} is increased of 0.001 and go back to step 2
   c. If marginal HR > marginal HR desired , β_{treat} is decreased of 0.001 and go back to step 2

These steps are iterated till β_{treat} required is found, with an approximation on the third decimals place. Three scenarios are simulated considering respectively three values for the marginal HR:

- Scenario a: HR=1
- Scenario b: HR=1.5
- Scenario c: HR=2

In each scenario 10,000 datasets consisting of 1,000 subjects are simulated.

For each dataset, the propensity score is estimated using a logistic regression model on the twelve confounders generated, then the hazard ratio and its standard error are measured using a robust Cox model to which the following propensity score based methods are applied:

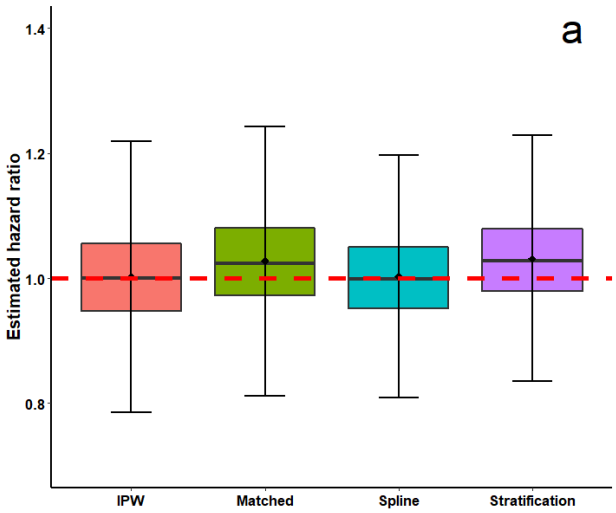
1. IPW
2. Greedy nearest-neighbour matching on PS, with caliper equal to 0.2 standard deviation of the logit(PS)
3. Covariate adjustment using the propensity score with spline transformation
4. Stratification on the propensity score using quintiles of the distribution

Once the estimates of the cumulative hazard ratios from the *i-th* simulation (HR_i) are obtained, their distribution and the distribution of the differences with the true HR on the logarithm scale (i.e. $\log(HR_i) - \log(HR_{true})$) were represented using boxplots.

The ratio between the mean standard error and the standard deviation of the estimated log-hazard ratios across the 10,000 simulated datasets was measured; it indicates whether, for a given estimation method, the estimated standard error of the estimated treatment effect is correctly estimating the sampling variability of the treatment. The 95% confidence interval for each HR_i is computed and the proportion of 95% confidence intervals that covered the true hazard ratio (coverage rate) is determined. Finally, the Mean Squared Error was calculated on the logarithm scale as: $(\frac{1}{10000}) \sum_{i=1}^{10000} (\log(HR_i) - \log(HR_{true}))^2$. The proportion of subjects treated is fixed on 0.38 and the average number of matched pairs formed across the 10,000 simulated samples is equal to 310 for all three scenarios. Thus, the 81.6% of treated subjects is approximately matched with an untreated subject. The average censorship rate is 16.6%, 13.9% and 12.8% for the three scenarios respectively.

2.2. Results

The distribution of the estimated hazard ratios shows a precise estimation of the hazard ratio in the scenario a, but as the true marginal hazard ratio increases, there is a tendency to overestimate the effect by all the methods except for the IPW estimator (figure 6).



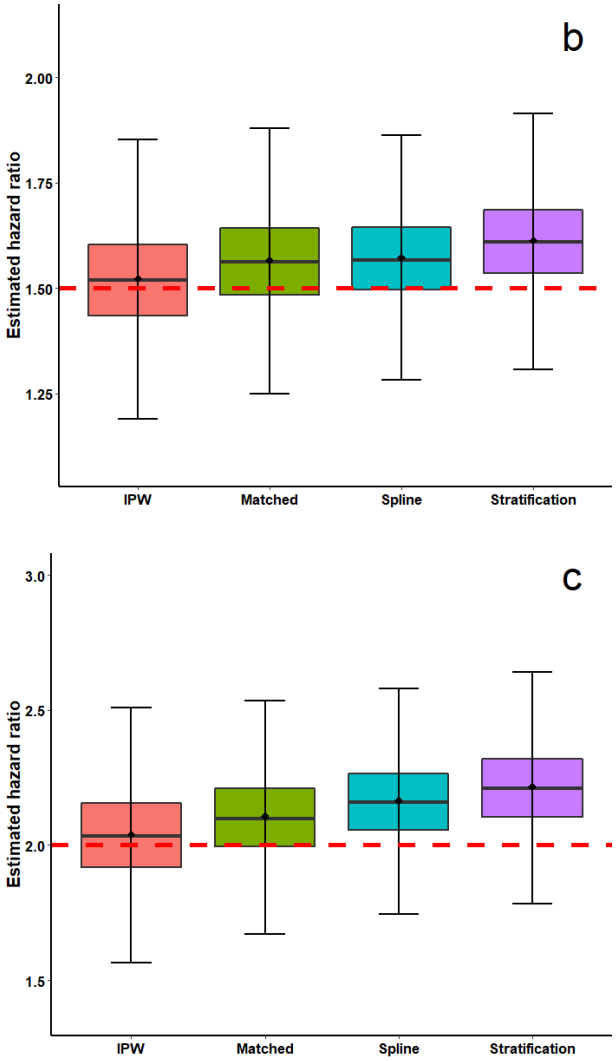


Figure 6. Distribution of the estimated hazard ratios in each simulation in the scenario a, b and c. Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

The bias is close to zero for the IPW and for propensity score as covariate with spline transformation in the scenario a, but it tends to assume higher values in scenarios b and c in every method. The only one with minimal bias in all of the three scenarios is the IPW (figure 7).

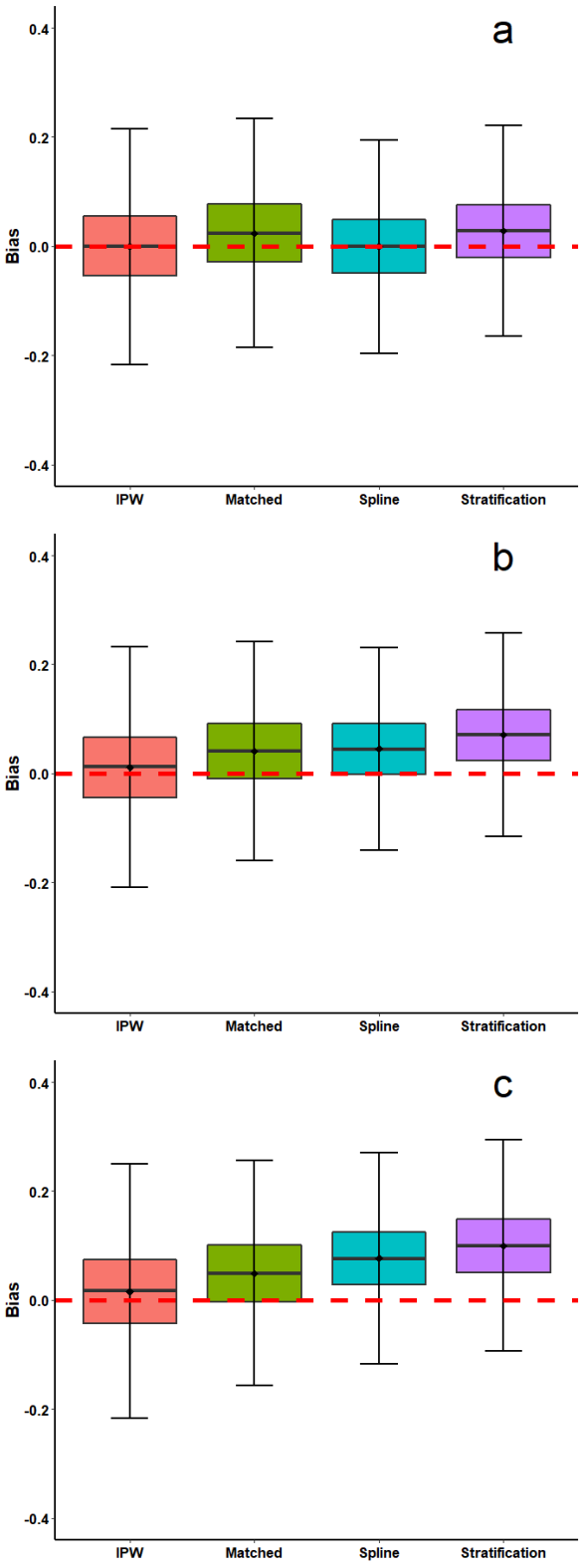


Figure 7. Distribution of the differences between the estimated log(HR) in each simulation and the true log(HR) in the scenario a, b and c. The horizontal solid lines represent the bias (mean of the differences). Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

The mean standard error and the standard deviation of the estimated log-HR are reported in the following table, together with their ratios:

		IPW	Matching	Spline	Stratification
Scenario a					
	Mean(SE)	0.090	0.087	0.078	0.078
	SD	0.081	0.078	0.072	0.072
	Mean(SE)/SD	1.113	1.110	1.094	1.082
Scenario b					
	Mean(SE)	0.091	0.086	0.077	0.076
	SD	0.083	0.076	0.070	0.070
	Mean(SE)/SD	1.100	1.124	1.092	1.078
Scenario c					
	Mean(SE)	0.095	0.087	0.077	0.076
	SD	0.089	0.078	0.073	0.073
	Mean(SE)/SD	1.073	1.114	1.054	1.040

Table 1. Mean Standard Error (SE), Standard Deviation (SD) and their ratio for each method in the three scenarios

In the three scenarios the average of the ratios between the mean standard error and the standard deviation of the estimated log-HR is close to one for every method (figure 8).

These results indicate that the average standard error is precise in estimating the sample dispersion of the estimated log-hazard ratios. In the scenario a and b, the averages of the ratios are very similar and close to the value 1.1 In the scenario c these values tend to be closer to 1 for all methods except matching on PS.

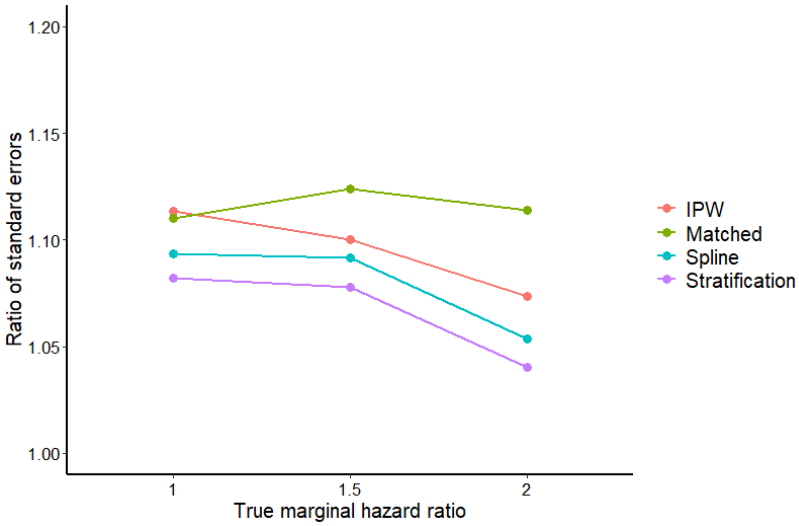


Figure 8. Ratio of mean standard error to standard deviation of estimated log-hazard ratios. The three scenarios (a, b and c) are here displayed on the horizontal axis (true marginal HR 1, 1.5 and 2). Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

Coverage rate of 95% confidence intervals is greater than 0.95 in all the scenarios for IPW. For matching on propensity score, it is greater than 0.95 in the scenario a, while it is equal to or less than 0.95 in the scenario b and c, respectively. For the other methods, the coverage rate is greater than 0.95 in the first scenario and it strongly decreases to very low values at the increasing of the true marginal HR (figure 9).

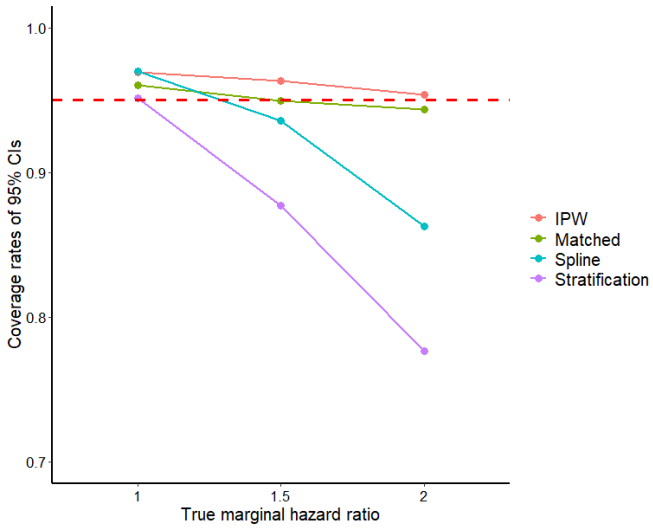


Figure 9. Coverage rates of 95% confidence intervals (CIs) for the estimated hazard ratios. The three scenarios (a, b and c) are here displayed on the horizontal axis (true marginal HR 1, 1.5 and 2). Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

In the end, figure 10 presents the MSE of the estimated treatment effects. The MSE values are all close to zero and increase together with the true marginal hazard ratio, even if this increase is very slight for all the four methods. The IPW estimator turns out to have the lowest MSE in every scenario, followed by the matched estimator. Stratification on PS is the methods with the highest values of MSE in all the three scenarios.

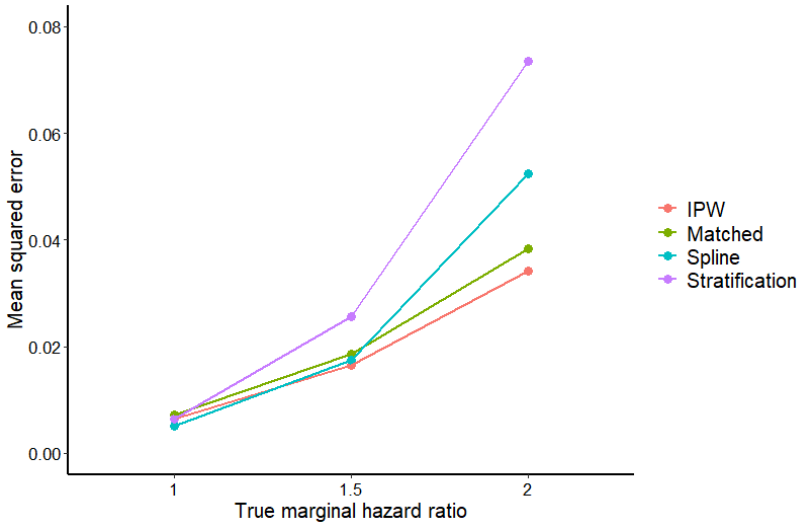


Figure 10. Mean squared error of the estimated log-hazard ratio. The three scenarios (a, b and c) are here displayed on the horizontal axis (true marginal HR 1, 1.5 and 2). Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

3. Simulations on weighted all-cause hazard ratio

3.1. Simulation protocol

Data are generated in order to represent a more simplified reality than that considered in the simulations on the standard all-cause-hazard ratio, since the weighted all-cause hazard ratio is more complex to estimate than the unweighted one.

Three baseline confounders (X_1-X_3) with different distributions (Normal, Binomial and Poisson) were generated.

- $x_1 \sim Bin(1; 0.7)$
- $x_2 \sim Poisson(1.3)$
- $x_3 \sim N(75; 10)$

For each subject, the probability of being assigned to one of two treatments is determined from the following logistic model:

$$\boldsymbol{l}ogit(p_i) = \alpha_{0,treat} + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \quad (43)$$

The intercept of the model ($\alpha_{0,treat}$) is set equal to 0.1 in order to obtain the desired proportion of treated subjects equal to 38%. The regression coefficients $\alpha_1 - \alpha_3$ have been set as follows:

- $\alpha_1 = 0.53$
- $\alpha_2 = 0.38$
- $\alpha_3 = -0.01$

The treatment status is generated from a Bernoulli distribution with subject-specific parameter p_i and the outcome is generated using a data-generating process for time to-event outcomes described by Bender et al [9]. A linear predictor is defined for the *i-th* subject as:

$$LP_1 = \beta_{treat1} Z + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (44)$$

$$LP_2 = \beta_{treat2} Z + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (45)$$

A random number is generated from a standard uniform distribution $u_1 \sim U(0,1)$ and the censorship are generated from the following uniform distribution $u_2 \sim U(0.5,2)$. The regression coefficients $\beta_1 - \beta_3$ have been set as follows:

- $\beta_1 = 0.19$
- $\beta_2 = 0.22$
- $\beta_3 = -0.01$

Two different endpoints are generated: death (event=1) and recurrence (event=2).

The event times for each subjects are generated from a Weibull distribution as follows:

- $T_1 = -\log(u)/(ke^{LP_1})^{1/p}$, $with\ k = 0.5\ and\ p = 1$ (46)
- $T_2 = -\log(u)/(le^{LP_2})^{1/q}$, $with\ l = 0.7\ and\ q = 2$ (47)

The corresponding hazard functions are:

- $\lambda_1(t) = \lambda_{01}(t)e^{LP_1}$ $with\ \lambda_{01}(t) = pk^p t^{p-1}$ (48)
- $\lambda_2(t) = \lambda_{02}(t)e^{LP_2}$ $with\ \lambda_{02}(t) = pl^q t^{q-1}$ (49)

Setting the parameters in this way, the assumption of equal baseline hazards across the components of the composite endpoint within each group of treatment made by Ozga et al [9] is violated. This was done because we aim to check whether the robustness of the estimator shown by Ozga is preserved even when estimating a marginal weighted all-cause-hazard-ratio with PS-based methods. However, just in order to ease the evaluation of the performance of the methods, another assumption was made: the treatment effect is not time dependent and thus it is represented by a single number.

This data-generating process results in a conditional treatment effect with a hazard ratio represented by $e^{\beta_{treat1}}$ for the endpoint 1 and $e^{\beta_{treat2}}$ for the endpoint 2.

The process to generate data with a chosen marginal cumulative weighted hazard ratio is similar to that described for the unweighted HR simulation protocol (Chapter 2), except for the fact that in this case two parameters, i.e. β_{treat1} and β_{treat2} have to be found. However, the iterative process can just be focused on searching for the value of β_{treat2}, leaving β_{treat1} fixed.

In the simulations, the weighted all-cause hazard ratio is estimated at a predefined time-point (i.e. at time 1) for simplicity. This can be done thanks to the constant HR assumption mentioned above.

Nine scenarios are simulated considering three different values for the marginal HR and three different types of weights:

	$(w_1;w_2)=(1;1)$	**$(w_1;w_2)=(1;0.5)$**	**$(w_1;w_2)=(1;0.8)$**
Scenario a	HR=1	HR=1	HR=1
Scenario b	HR=1.5	HR=1.5	HR=1.5
Scenario c	HR=2	HR=2	HR=2

Table 2. Scenarios of the simulations on weighted all-cause hazard ratio

In each scenario, we simulated 1,000 datasets, each consisting of 1,000 subjects.

The propensity score is estimated using a logistic regression model on the 3 confounders generated, then the ratio of the cumulative hazard and its standard error are measured using the non-

parametric estimator [9] to which two propensity score based methods are applied: IPW and greedy nearest-neighbour matching on PS.

Once the estimates of the cumulative hazard ratios from the *i-th* simulation (HR_i) are obtained, their distribution and the distribution of the differences with the true HR on the logarithm scale (i.e. $\log(HR_i) - \log(HR_{true})$) were represented using boxplots. The ratio between the mean standard error and the standard deviation of the estimated log-hazard ratios across the 1,000 simulated datasets was measured; it indicates whether, for a given estimation method, the estimated standard error of the estimated treatment effect is correctly estimating the sampling variability of the treatment. The 95% confidence interval for each HR_i is computed and the proportion of 95% confidence intervals that covered the true hazard ratio (coverage rate) is determined. Finally, the Mean Squared Error was calculated on the logarithm scale as: $(\frac{1}{1000}) \sum_{i=1}^{1000} (\log(HR_i) - \log(HR_{true}))^2$. The proportion of subjects treated is fixed on 0.58 and the average number of matched pairs formed across the 1,000 simulated samples is 408 in all scenarios. Thus, the 70.3% of treated subjects is approximately matched with an untreated subject. The average censorship rate for each scenario is reported in the following table:

	$(w_1;w_2)=(1;1)$	**$(w_1;w_2)=(1;0.5)$**	**$(w_1;w_2)=(1;0.8)$**
Scenario a	29.3%	29.4%	29.4%
Scenario b	23.3%	22.7%	25.4%
Scenario c	19.0%	19.4%	19.7%

Table 3. Censorship rate for each scenario of the simulations on weighted all-cause hazard ratio

3.2. Results

The distribution of the estimated cumulative hazard ratios shows that the estimates obtained through the IPW estimator are more precise that those obtained through the matched sample. Moreover, considering the two methods separately, the distributions in the nine scenarios are very similar (figure 11).

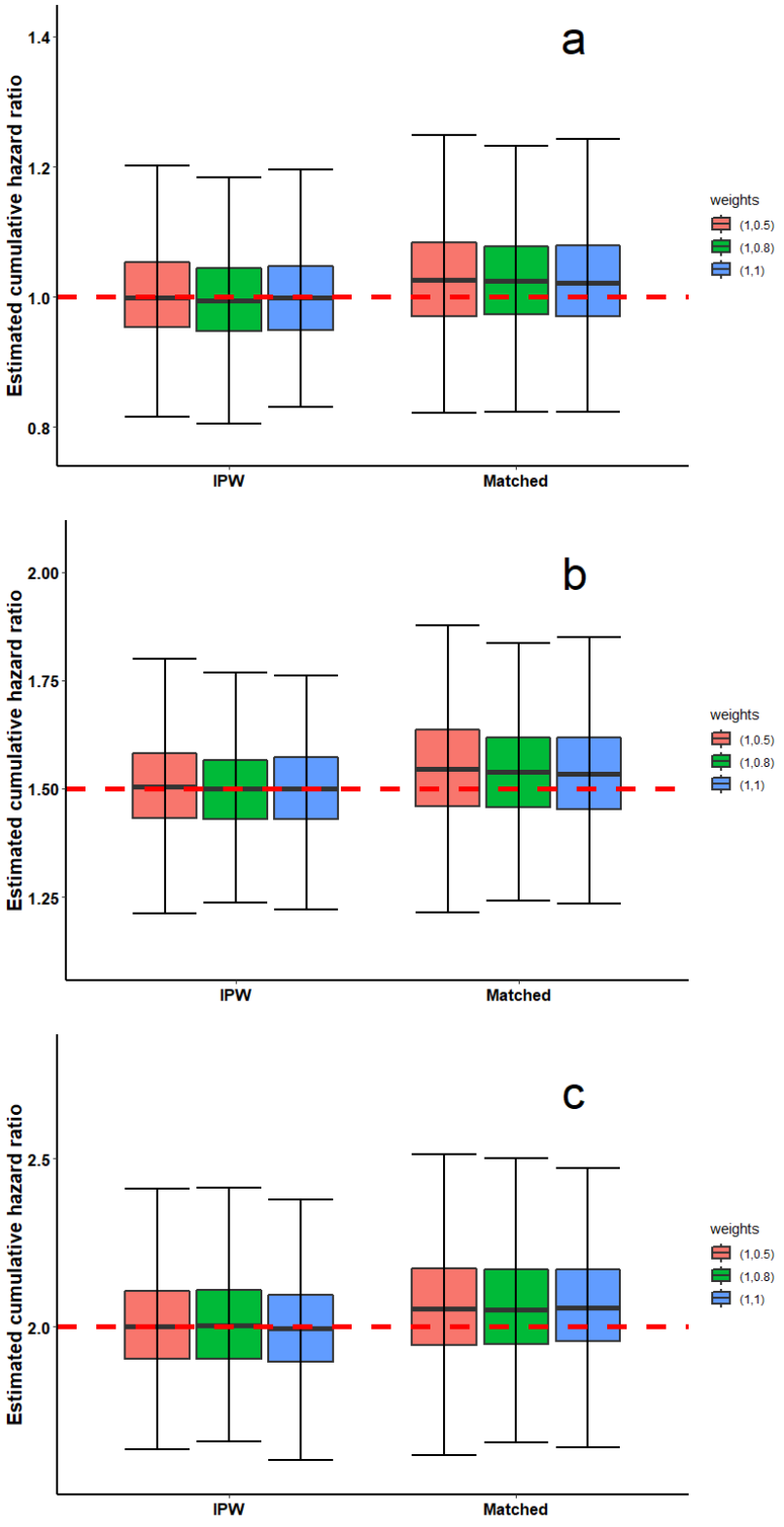


Figure 11. Distribution of the estimated hazard ratios in all scenarios. Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

The bias is close to zero for the IPW method in all scenarios, but it tends to assume higher values for the propensity score matching (figure 12).

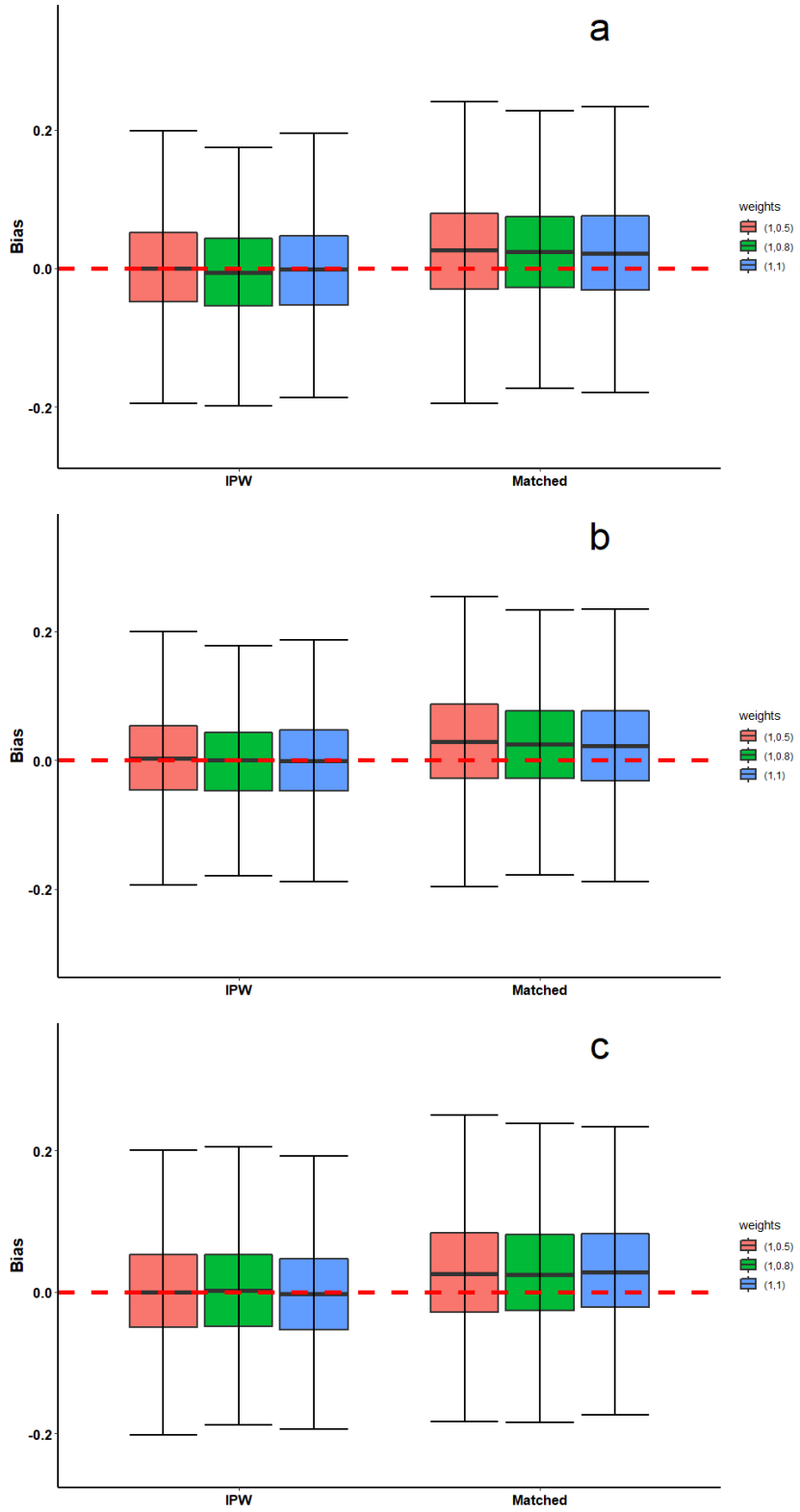


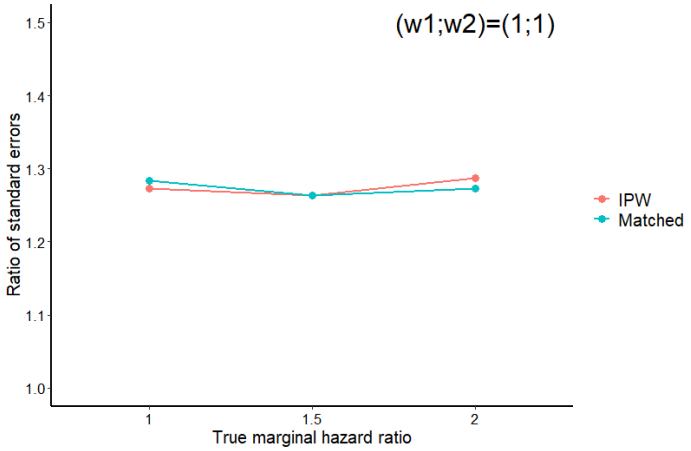
Figure 12. Distribution of the differences between the estimated log(HR) in each simulation and the true log(HR) in the scenario a, b and c. The horizontal solid lines represent the bias (mean of the differences). Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

The mean standard error and the standard deviation of the estimated log-HR together with their ratio are reported in the following table:

	Scenario a			Scenario b			Scenario c		
	(1;1)	**(1;0.5)**	**(1;0.8)**	**(1;1)**	**(1;0.5)**	**(1;0.8)**	**(1;1)**	**(1;0.5)**	**(1;0.8)**
IPW									
Mean(SE)	0.089	0.095	0.090	0.089	0.095	0.090	0.091	0.099	0.092
SD	0.070	0.075	0.070	0.071	0.074	0.068	0.071	0.076	0.072
Mean(SE)/ SD	1.273	1.279	1.291	1.264	1.285	1.317	1.287	1.297	1.285
Matching									
Mean(SE)	0.101	0.107	0.101	0.099	1.105	0.099	0.100	0.108	0.102
SD	0.079	0.083	0.077	0.079	0.085	0.078	0.079	0.086	0.083
Mean(SE)/ SD	1.284	1.299	1.312	1.263	1.228	1.271	1.274	1.258	1.230

Table 4. Mean Standard Error (SE), Standard Deviation (SD) and their ratio for each method in all scenarios

The average of the ratios between the mean standard error and the standard deviation of the estimated log-hazard ratios tends to be larger than one in both methods for all the three scenarios. It means that the standard error tends to overestimate the sample variability (figure 13).



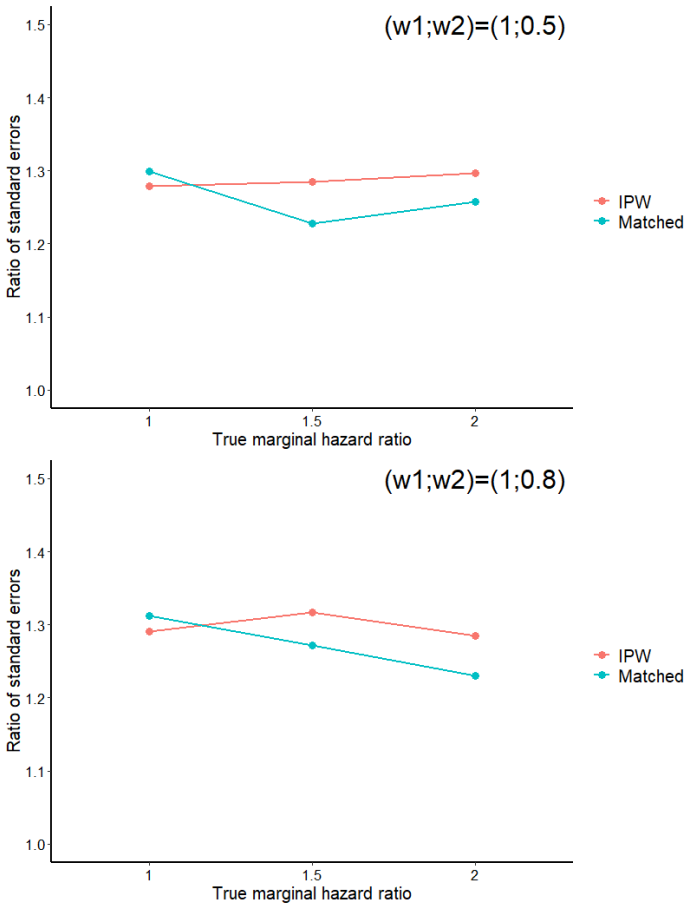
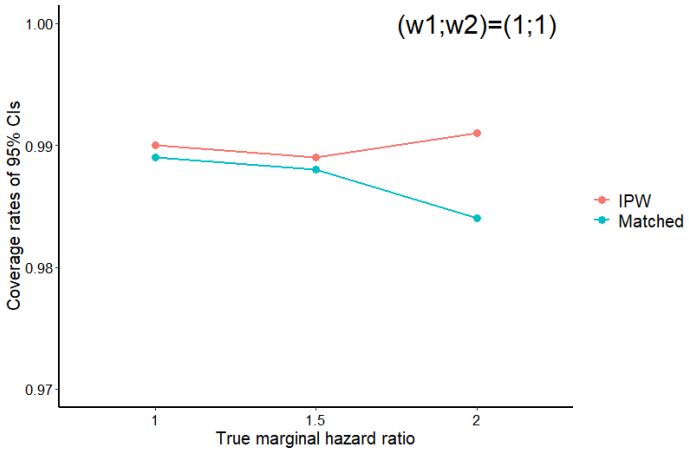


Figure 13. Ratio of mean standard error to standard deviation of estimated log-hazard ratios in all scenarios. The three scenarios (a, b and c) are here displayed on the horizontal axis (true marginal HR 1, 1.5 and 2). Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

Coverage rates of 95% confidence intervals is optimal in all scenarios for both methods (figure 14).



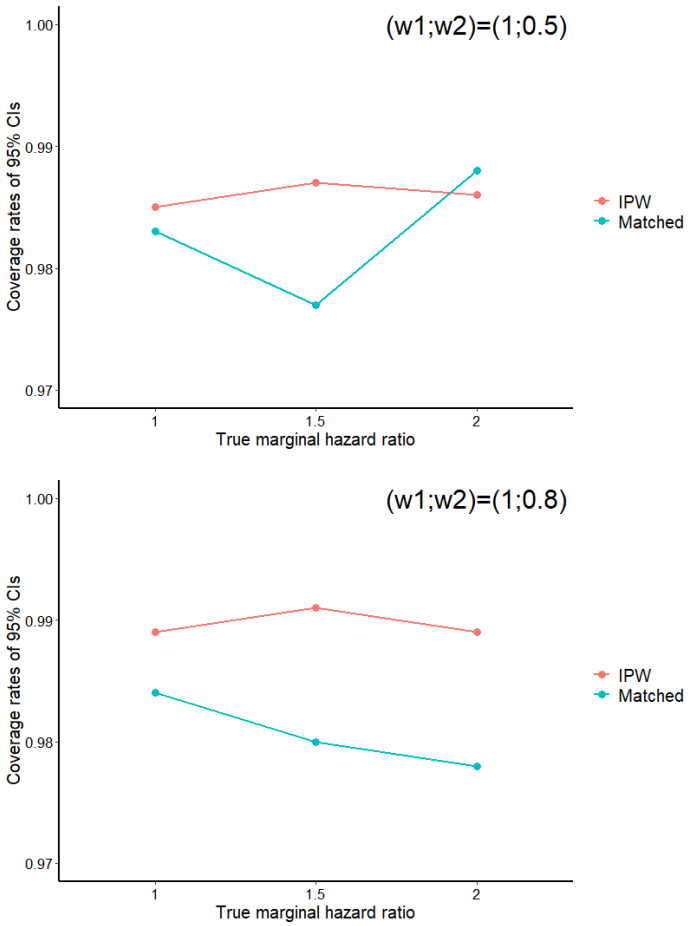
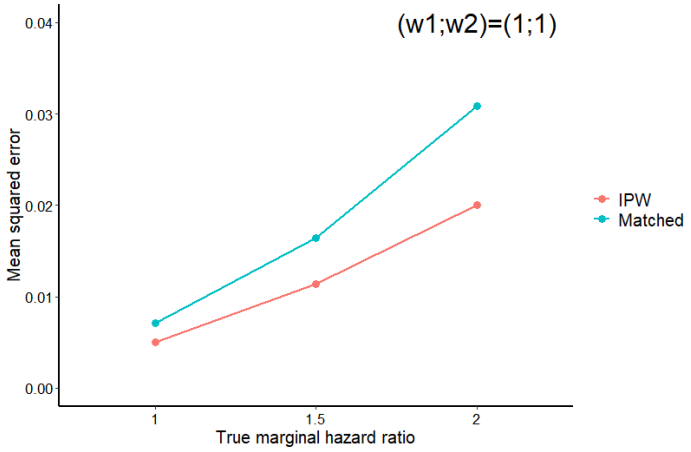


Figure 14. Coverage rates of 95% confidence intervals (CIs) for the estimated hazard ratios. The three scenarios (a, b and c) are here displayed on the horizontal axis (true marginal HR 1, 1.5 and 2). Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

In the end, figure 15 presents the MSE of the estimated treatment effects. The MSE increases together with the true marginal hazard ratio, even if this increase is very slight for both methods. The IPW estimator turns out to have the lowest MSE in every scenario.



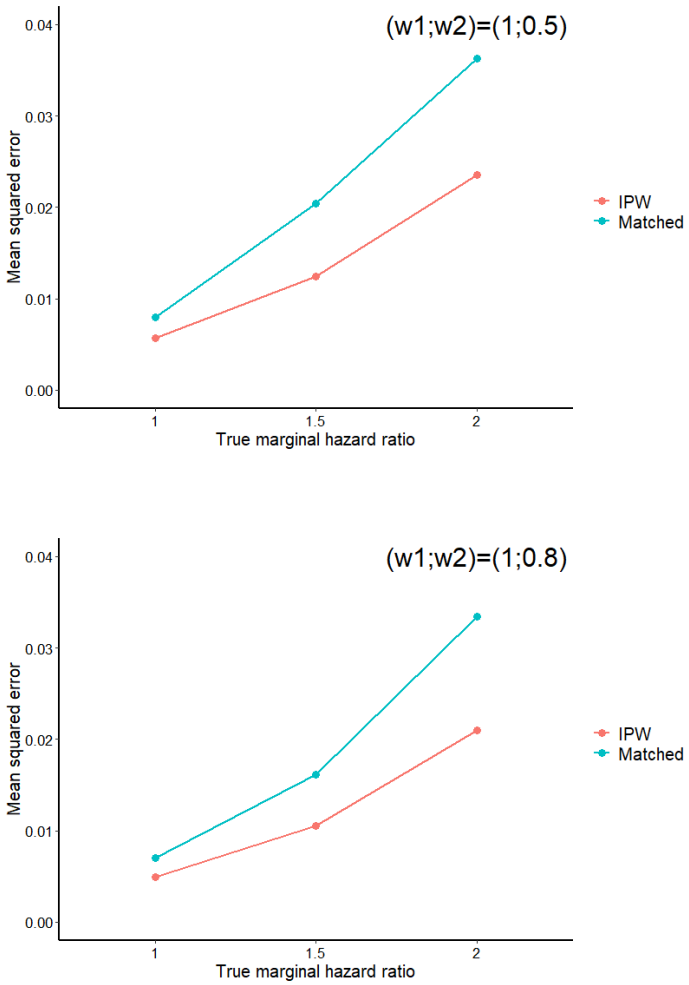


Figure 15. Mean squared error of the estimated log-hazard ratio. The three scenarios (a, b and c) are here displayed on the horizontal axis (true marginal HR 1, 1.5 and 2). Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles

4. Application

4.1. The clinical context

HCC is a major health problem, as in 2016 one million incident cases of liver cancer globally and 829,000 deaths were recorded. It ranks as the fifth most common cause of cancer in men and the seventh in women representing a third of all cancer-related deaths and the leading cause of death in patients with liver cirrhosis [21].

The HCC is more common in East Asia, however its incidence is increasing in the Western World.

Hepatic resection is the first-line therapeutic option and it is accepted as a safe treatment with a proven impact on prognosis, with a low operative mortality as the result of advances in surgical techniques and perioperative management. Nevertheless, surgical resection is applicable in only about 20% to 30% of patients with HCC, since most have poor hepatic reserve function caused by underlying chronic liver disease and multifocal hepatic distributions of HCC.

Hepatic resection is one of the curative treatments for hepatocellular carcinoma, however the recurrence rate of HCC even after curative resection is quite high, estimated to be approximately 50% during the first three years and more than 70% during the first five years after curative resection, and so the postoperative long term results remain unsatisfactory.

Although surgical treatment has been adopted in the last years in more patients outside the guidelines with satisfactory results in term of mortality, morbidity and short term oncological outcomes, the limits of this approach remain the long term disease free survival [8]. Moreover, the comparison between the efficacy of wedge resection and anatomic resection, the two most common types of laparoscopic hepatectomy performed, is an open field of investigation. The WR is a surgical procedure to remove a triangle-shaped slice of tissue, usually used to remove a tumor or some other type of tissue that requires removal and typically includes a small amount of normal tissue around it. The AR is defined as the complete removal of at least one liver's segment containing the tumor together with the related portal vein and the corresponding hepatic territory.

4.2. The HERCOLES study

The HERCOLES Project (Hepatocarcinoma Recurrence on the Liver Study) is an Italian observational multicentric study on surgical treatment and survival endpoints of patients affected by hepatocellular carcinoma (HCC) [8].

Data are collected prospectively and anonymized prior to the analysis.

The primary aim of the study is to evaluate the impact of surgical resection, anatomic and wedge, on Disease-Free-Survival, Overall Survival and Tumor-Specific-Survival within a national framework; the secondary aim is to evaluate the role of different clinical, biochemical, radiological and histopathological variables in determining the post-surgery recurrence.

The inclusion criteria are the following:

- No age limit;
- Hepatocarcinoma diagnosis confirmed at histological specimen;
- Patient with a first diagnosis of HCC, or with a recurrence/persistence treated with surgical resection at the participating centers.

On the other hand, the exclusion criteria are:

- Surgical resection performed as down-staging therapy towards transplantation;
- Patients treated with surgery for non-curative purposes (palliation, best supportive care, etc.);
- Primary mixed etiology tumors (i.e. hepatocolangiocarcinoma);
- Patients with other previous cancers.

For the purpose of this thesis, we considered data of patients enrolled between 2008 and 2017. The Italian centers which joined the project are reported in table 5.

Center name	**City**	**Overall (n=1089) n(%)**	**Anatomic resection (n=722) n(%)**	**Wedge resection (n=367) n(%)**
IRCCS Ospedale San Raffaele	Milano	540 (49.6)	442 (61.2)	98 (26.7)
Fondazione IRCCS Istituto Nazionale dei Tumori	Milano	121 (11.1)	47 (6.5)	74 (20.2)
Ospedale San Gerardo	Monza	85 (7.8)	21 (2.9)	64 (17.4)
Fondazione IRCCS Policlinico San Matteo	Pavia	78 (7.2)	54 (7.5)	24 (6.5)
Azienda Ospedaliera Spedali Civili di Brescia	Brescia	68 (6.2)	37 (5.1)	31 (8.4)

Ospedale Pierantoni Morgagni	Forlì	43 (3.9)	23 (3.2)	20 (5.4)
Policlinico Borgo Nuovo	Verona	43 (3.9)	32 (4.4)	11 (3.0)
Policlinico di Monza	Monza	29 (2.7)	17 (2.4)	12 (3.3)
Istituto Fondazione Poliambulanza	Brescia	26 (2.4)	16 (2.2)	10 (2.7)
Ospedale Maggiore	Crema	20 (1.8)	20 (2.8)	0 (0.0)
Policlinico di Bari Ospedale "Giovanni XXIII"	Bari	17 (1.6)	6 (0.8)	11 (3.0)
Chirurgia Oncologica Epatobiliopancreatica	Parma	8 (0.7)	4 (0.6)	4 (1.1)
Ospedale San Paolo	Savona	6 (0.6)	2 (0.3)	4 (1.1)
Ospedale Sacco	Milano	5 (0.5)	1 (0.1)	4 (1.1)

Table 5. The Italian centers that joined the study

4.3. Statistical analysis

As mentioned in the previous section, one of the aims of the HERCOLES study is to compare the impact of anatomic vs wedge surgical resection to treat liver cancer. The main endpoint of interest is the disease free survival (DFS) defined as time from treatment to the first of the following cause-specific events: local recurrence (appearing on the surgical margin), other hepatic recurrence and death without recurrence.

The purpose of this study is to quantify the marginal treatment effect of surgical resection on DFS through PS-based methods. In other words, the aim is to estimate the impact of treatment while trying to remove or reduce the effect of the different distribution of some prognostic factors (confounders) between the two treatment groups.

We recall that, loosely speaking, a confounder is a variable that influences both the outcome and the assignment of treatment, causing a spurious association.

Confounding factors are here identified using a logistic model with Least Absolute Shrinkage and Selection Operator (LASSO) penalty [22], considering variables associated with the outcome of interest from clinical knowledge.

The type of surgical resection is entered as dependent variable, Y, in the logistic regression model and is coded as 0 for anatomic resection and 1 for wedge resection. The probability of being treated with wedge resection given the covariates x_i is calculated as follows:

$$P(Y = 1|x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}} \quad (50)$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ are covariates of the *ith* observation and include the variables shown in tables 7 and 8. The parameter β_0 is the intercept and β_j (j=1,…,k) is the coefficient corresponding to the *jth* covariate. The logistic LASSO estimator $\widehat{\beta_0}, \ldots, \widehat{\beta_k}$ is defined as the minimizer of the negative log likelihood:

$$\sum_{i=1}^{n}[-y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) + \log(1 + e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}})] \quad (51)$$

subject to $\sum_{j=1}^{k}|\beta_j| \leq \lambda$. Here, λ>0 is a tuning parameter that controls the sparsity of the estimator (i.e., the number of coefficients with a value of zero) and is selected by cross-validation. We used the "glmnet" package in R to apply the logistic LASSO estimator on our data. The propensity score is estimated using a logistic regression model considering the type of surgical resection as the outcome and the confounders selected by the LASSO model as covariates. The marginal treatment effect is measured in terms of marginal hazard ratio. For the case of the standard unweighted composite end-point (i.e. DFS), this is estimated by applying the PS-based methods to a Cox model. To account for the different clinical relevance of each specific event (death obviously corresponds to the worst event, but also local recurrence is considered more severe than other recurrence) the non-parametric estimator proposed by Ozga and Rauch [9] for a weighted cumulative hazard ratio was considered, again in combination with the application of PS-based methods to obtain a marginal measure. The PS-based methods used to estimate a marginal effect of the type of surgical resection on DFS are the following:

I. Inverse probability weighting (IPW)
II. Greedy nearest-neighbour matching on PS
III. Covariate adjustment using the propensity score with spline transformation
IV. Stratification on the propensity score using the quintile of the distribution

4.4. Results

A total of 1089 patients were enrolled. The 34% (n=367) of the total underwent wedge-type surgery and the remaining 66% (n=722) anatomic-type surgery.

The patients alive and without any type of recurrence at the end of the follow-up are 524 (48%); 91 (8%) has developed a local recurrence; 408 (38%) has other types of recurrence and the remaining 66 (6%) died.

The description of demographical and baseline clinical characteristics are reported overall and by type of surgery (table 7-8).

The distribution of the three endpoints is different between the two treatment groups, while the median of follow-up is similar in the two groups (table 6).

Type of event	**Overall (n=1089) n(%)**	**Anatomic resection (n=722) n(%)**	**Wedge resection (n=367) n(%)**	**p-value**
Local recurrence	91 (8.4)	45 (6.2)	46 (12.5)	<0.001
Non-local recurrence	408 (37.5)	282 (39.1)	126 (34.3)	
Death	66 (6.1)	31 (4.3)	35 (9.5)	
No event	524 (48.1)	364 (50.4)	160 (43.6)	
Median follow-up (months)	55	54	61	

Table 6. Distribution of the endpoints and median follow-up divided by type of surgery. P-value obtained through Pearson's Chi-squared test.

The disease free survival curve stratified by type of surgery (figure 16) shows a higher survival probability for patients who have received an anatomical resection (p=0.005)

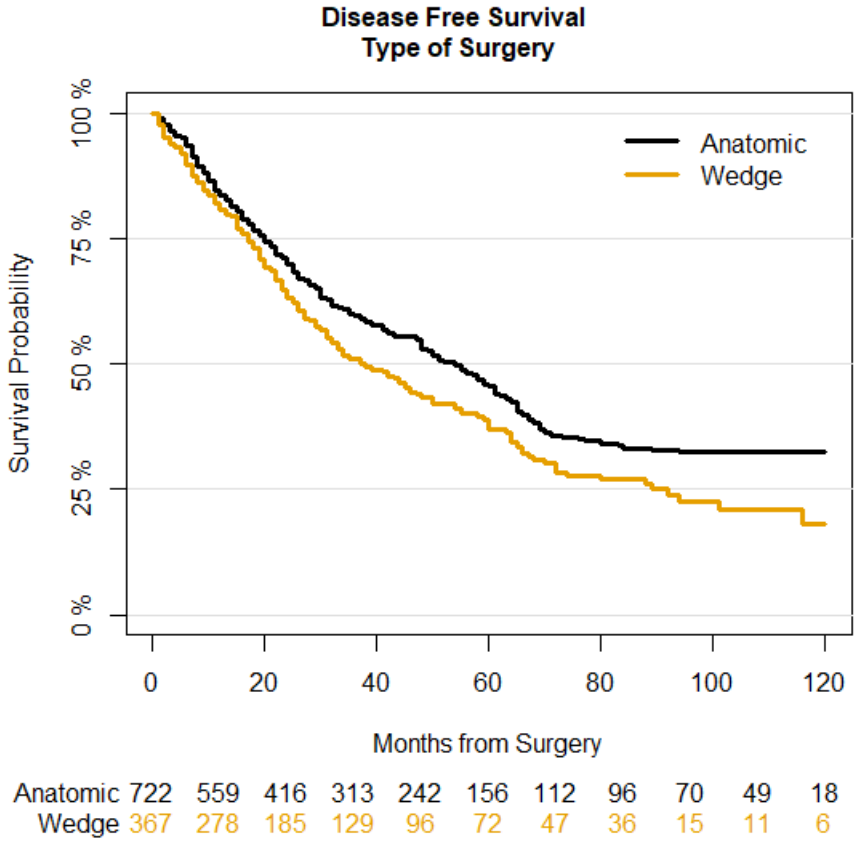


Figure 16. Disease Free Survival stratified for type of surgery

4.4.1. The confounding factors

The baseline characteristics available from the HERCOLES data and known in the clinical practice to be associated with both the choice of the treatment and the outcome are the following:

- Sex (M/F)
- Age (years)
- Cirrhosis (Y/N)
- Child Pugh Grade
- Hepatitis B virus (Y/N)
- Hepatitis C virus (Y/N)
- Alcoholic (Y/N)
- Number of nodules
- Histological grading
- Adjuvant therapy (Y/N)
- American Society of Anaesthesiologists (ASA)

- Microvascular Invasion (Y/N)
- International Normalized Ratio (INR)
- Platelets (thousands/µl)
- Larger nodule size (cm)
- Bilirubin (mg/dL)
- Surgical margin distance (mm)

Twelve of these confounders, described in tables 7-8, were selected using the LASSO model. Most of them are associated with the choice of the treatment, while age, HCV and adjuvant therapy do not show a significant association. Nevertheless, all of them are included in the propensity score estimation, starting from a clinical background.

Characteristics	**Overall (n=1089)**	**Anatomic resection (n=722)**	**Wedge resection (n=367)**	**p-value**
Age (years)				
Mean±SD	74.7±9.91	74.91±9.92	74.15±9.88	0.231
Median	77	77	75	
I-III quartile	69-81	70-81	68-82	
INR				
Mean±SD	1.2±0.23	1.20±0.23	1.15±0.22	<0.001
Median	1.1	1	1	
I-III quartile	1 -1.3	1-1.3	1 -1.2	
Platelets (thousands/µl)				
Mean±SD	182.7±88.35	194.59±89.89	159.25±80.34	<0.001
Median	172	183	148	
I-III quartile	116-232	125-250	95-193	
Larger nodule size (cm)				
Mean±SD	4.6±2.89	5±3.1	3±1.9	<0.001
Median	3.8	4	3	
I-III quartile	2.8-5	3-7	2-4	

Table 7. Continuous confounders. Abbreviations: INR=International Normalized Ratio; IQR=interquartile range. P-value obtained through two sample t-test.

Characteristics	Level	Overall (n=1089) n (%)	Anatomic resection (n=722) n (%)	Wedge resection (n=367) n (%)	p-value
Sex	M	815 (74.8)	536 (74.2)	279 (76.0)	0.571
	F	274 (25.2)	186 (25.8)	88 (24.0)	
Cirrhosis	No	335 (30.8)	246 (34.1)	89 (24.3)	0.001
	Yes	754 (69.2)	476 (65.9)	278 (75.7)	
Child Pugh Grade	A	985 (90.4)	641 (88.8)	344 (93.7)	0.012
	B	104 (9.6)	81 (11.2)	23 (6.3)	
HCV	No	497 (45.6)	323 (44.7)	174 (47.4)	0.439
	Yes	592 (54.4)	399 (55.3)	193 (52.6)	
Number of nodules	<2	862 (79.2)	595 (82.4)	267 (72.8)	<0.001
	≥2	227 (20.8)	127 (17.6)	100 (27.2)	
Histological grading	1	91 (8.4)	54 (7.5)	37 (10.1)	0.002
	2	748 (68.7)	521 (72.2)	227 (61.9)	
	3	250 (23.0)	147 (20.4)	103 (28.1)	
Adjuvant therapy	No	1068 (98.1)	711 (98.5)	357 (97.3)	0.259
	Yes	21 (1.9)	11 (1.5)	10 (2.7)	
ASA score	<3	603 (55.4)	428 (59.3)	175 (47.7)	<0.001
	≥3	486 (44.6)	294 (40.7)	192 (52.3)	

Table 8. Categorical confounders. Abbreviations: HCV=Hepatitis C virus; ASA=American Society of Anaesthesiologists. P-value obtained through Pearson's Chi-squared test.

The disease free survival stratified for the confounding factors are estimated in order to study the association between confounders and the outcomes (figure 17).

Cirrhotic patients have a higher probability to develop a recurrence or death compared to non-cirrhotic (p=0.008). The higher the ASA score, the number of nodules and the histological grade of the tumor (assessed through post-surgical biopsy), the higher the probability to have a recurrence or death (p<0.0001, p=0.004 and p<0.001, respectively). The Adjuvant therapy has a protective effect on recurrence or death (p=0.040). The Hepatitis C virus (HCV) and the Child Pugh Grade have no effect on the disease free survival (p=0.200 and p=0.400, respectively) but they are included in

the propensity score estimation because they have an impact on survival probability, according to clinical knowledge.

The larger nodule size and the age seem not to have an effect on the probability of having a recurrence or dying, while the International Normalized Ratio (INR) and the number of platelets (within the normal range) seem have a protective effect on the disease free survival (figure 18).

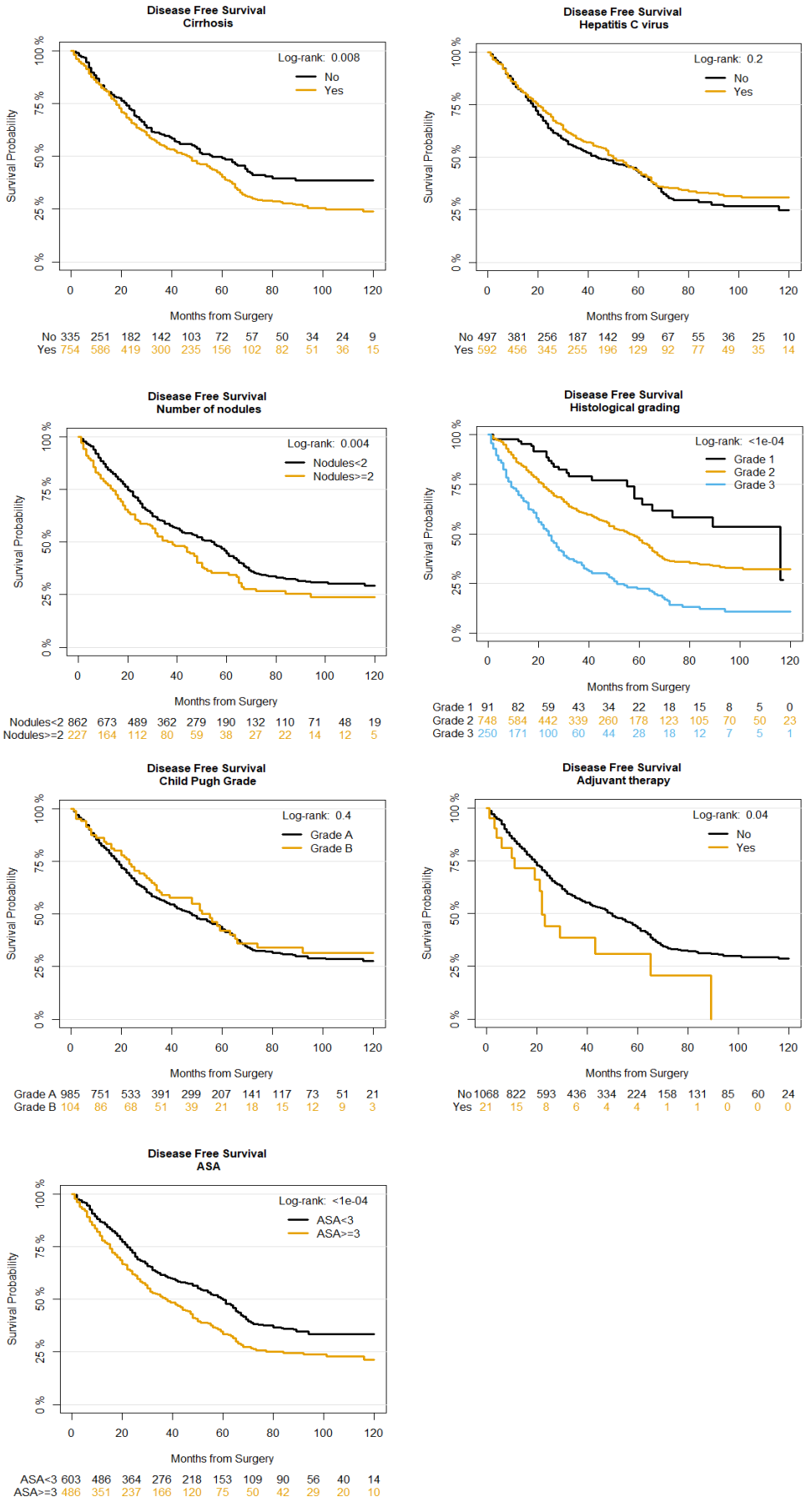


Figure 17. Disease free survival curve stratified for the categorical confounders

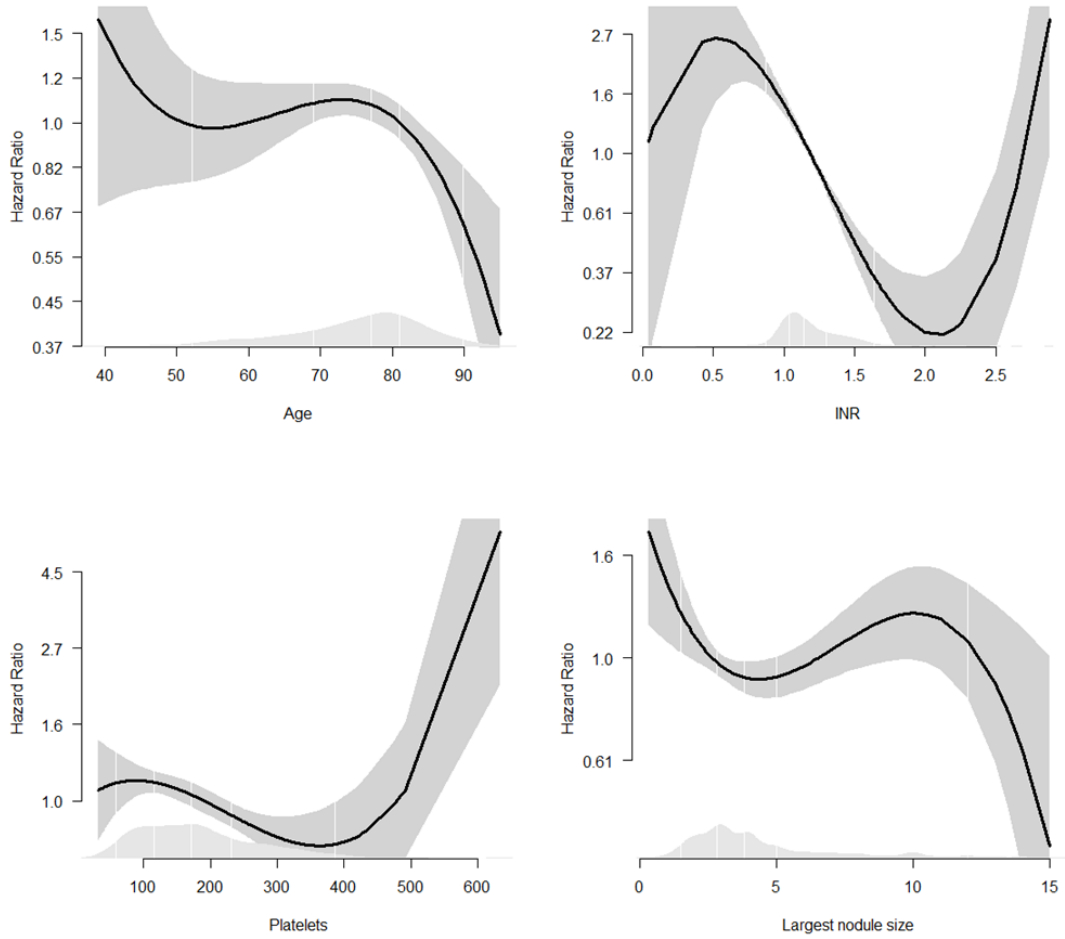


Figure 18. Smoothed HR distribution to study the effect of each continuous confounder on the disease free survival. Abbreviations: INR=International Normalized Ratio; HR=hazard ratio; SE=standard error; CI=confidence interval.

Propensity Score estimation

The PS distribution in the two groups of treatment (figure 19) shows that the PS well detects the differences between the two groups. In particular, as expected, the distribution of PS (i.e. probability of wedge) is generally higher for the patients that undergone the wedge resection (median: 0.46; I-III quartile: 0.33-0.58) than the group that received the anatomic surgical technique (median: 0.27; I-III quartile: 0.13-0.41).

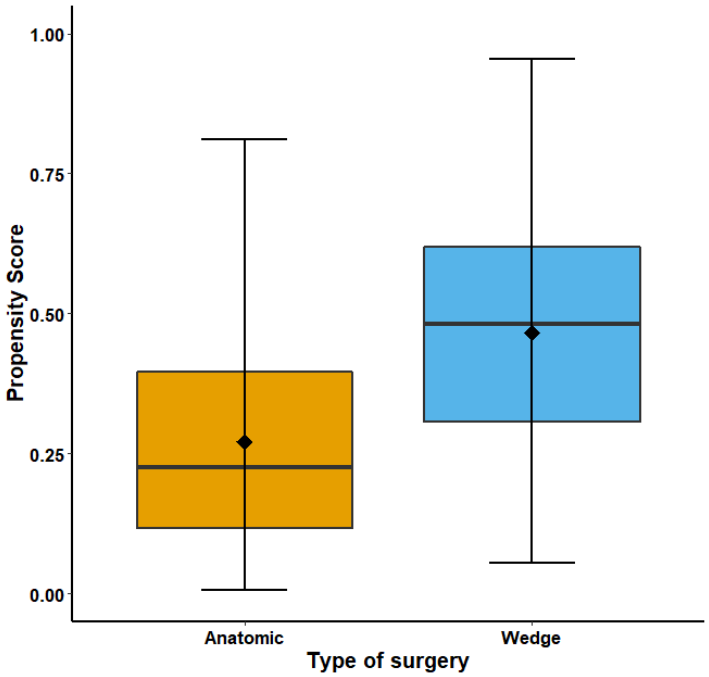


Figure 19. Distribution of the propensity score by type of surgery. Abbreviations: Anatomic = anatomic resection; Wedge = wedge resection.

Considering the coefficients of the PS model (table 9), only six confounders appear to be statistically associated with the choice of treatment: cirrhosis, INR, platelets, number of nodules, larger nodule size and the ASA score.

	OR	**95% CI**	**p-value**
Age	0.99	0.97-1.00	0.126
Cirrhosis	1.39	1.01-1.91	0.042
Child Pugh Grade	0.63	0.37-1.06	0.081
HCV	0.82	0.62-1.08	0.158
INR	0.45	0.24-0.85	0.014
Platelets	1.00	0.99-1.00	0.003
Number of nodules	1.73	1.24-2.41	0.001
Larger nodule size (cm)	0.72	0.67-0.78	<0.001
Histological grading (2 vs 1)	0.94	0.57-1.52	0.788
Histological grading (3 vs 1)	1.47	0.86-2.52	0.157
Adjuvant therapy	1.56	0.57-4.27	0.383
ASA score	1.55	1.16-2.07	0.003

Table 9. OR of the propensity score model with their confidence intervals and p-value. Abbreviations: HCV=Hepatitis C virus; INR=International Normalized Ratio; ASA=American Society of Anaesthesiologists

4.4.2. Results on marginal hazard ratio

Before the estimation of the ATE, the balancing between the two groups of treatment was evaluated for the IPW and PS matching methods. The number of matched pairs formed are 330 in total.

The absolute mean differences are lower than the conventional threshold of 0.01 for all the confounders, so a good balancing has been achieved (figure 20).

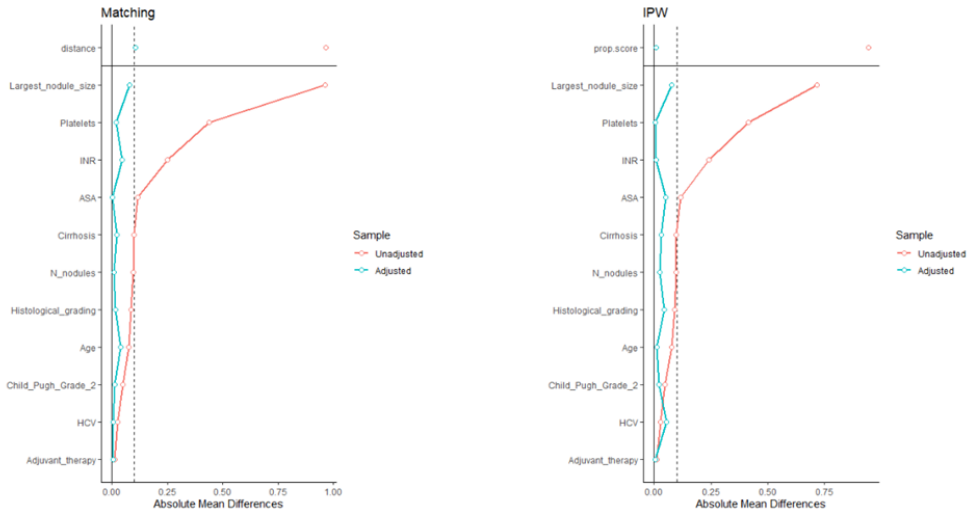


Figure 20. Covariate balance evaluation for the Inverse Probability Weighting (IPW) and propensity score matching

The estimate of the effect of the two types of surgical resections on the DFS (figure 21) proves that the choice of the surgical technique does not influence the standard composite endpoint. The proportional hazards assumption was checked on the adjusted model using the test based on Schoenfeld residuals (p= 0.354)

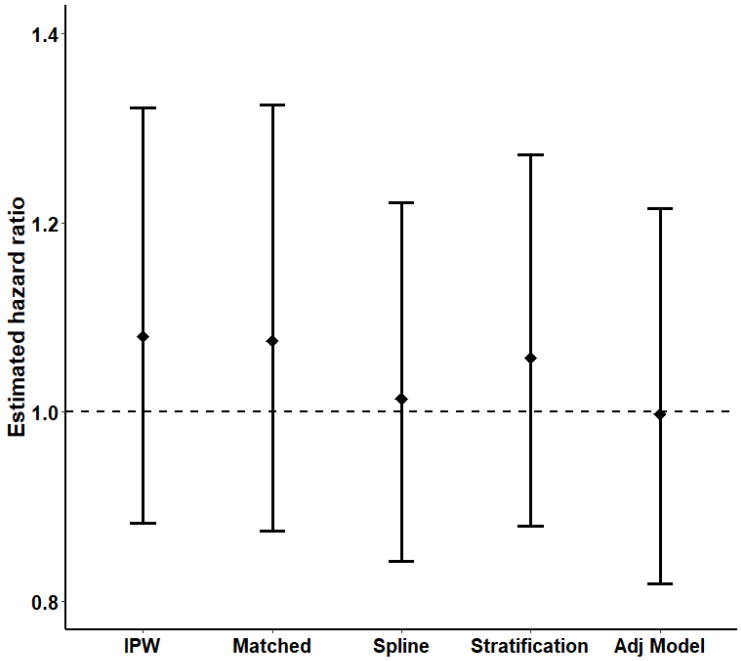


Figure 21. The estimated hazard ratios for the disease free survival and their 95% confidence intervals for each propensity score method. Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles; Adj Model=Adjusted Model with confounders as covariates.

Furthermore, it is of clinical interest to investigate the impact of the surgical resections on each cause-specific endpoint: local recurrence, non-local recurrence and death. To this purpose, a competing risk analysis is performed in order to consider each endpoint separately.

Considering the cumulative incidence of the three endpoints stratified for the two surgical resections (figure 22), the non-local recurrence has the highest incidence, while local recurrence and death have both a much lower incidence.

Focusing on every single endpoint, there is a statistically significant difference (Grey test) between the cumulative incidence of the two treatment groups for local recurrence ($p<0.001$) and death ($p<0.001$), in favour of anatomic resection. On the other hand, the control of other recurrence was in favour of wedge resection, although not statistically significant ($p= 0.188$)

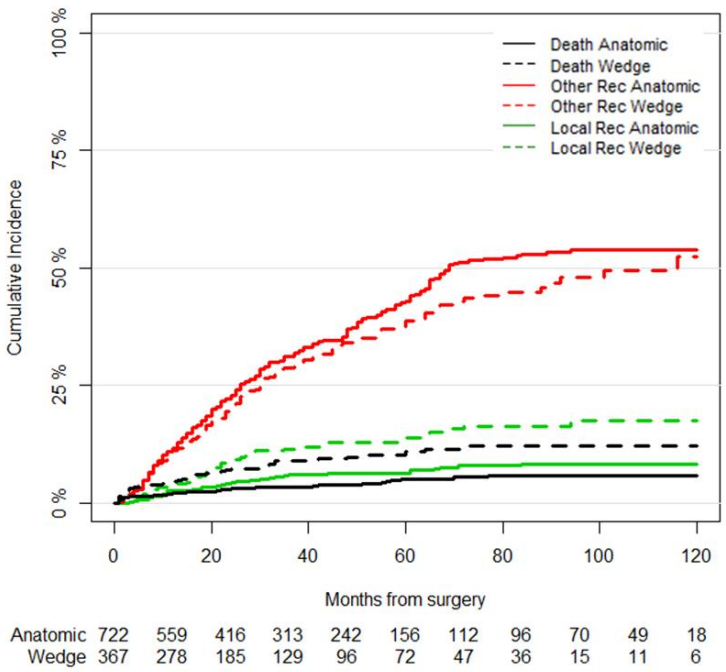


Figure 22. Cumulative Incidence of the cause-specific endpoints (local recurrence, non-local recurrence and death) stratified for type of surgery. Below the number at risk at different time points.

The results of the HR for each cause specific support different conclusion from those obtained in the analysis of the DFS endpoint (figure 23).

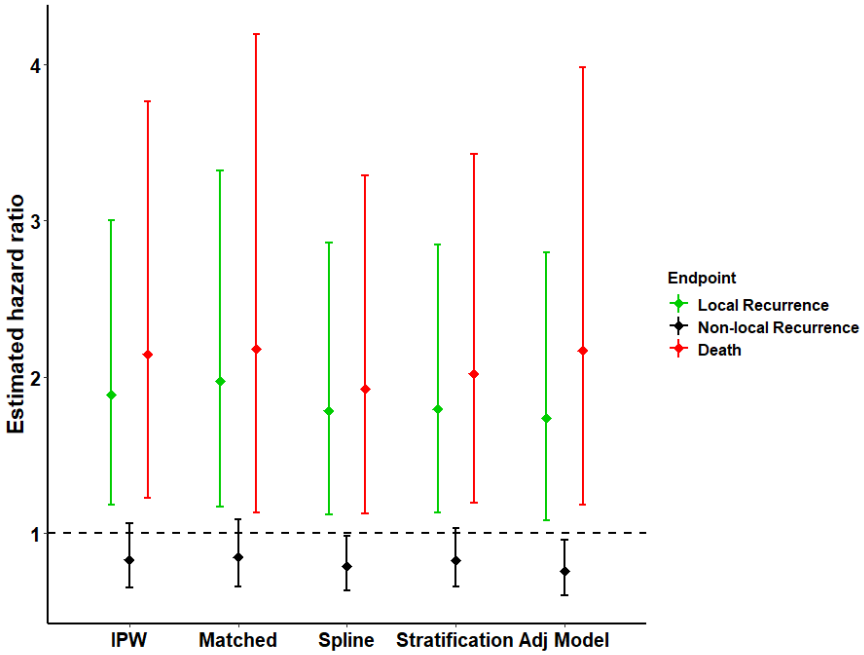


Figure 23. The adjusted hazard ratios for the cause-specific endpoints (local recurrence, non-local recurrence and death) and their 95% confidence interval for each propensity score method. Abbreviations: IPW=Inverse Probability Weighting; Matched=matching on propensity score; Spline= propensity score as covariate with spline transformation; Stratification= stratification by propensity score quintiles; Adj Model=Adjusted Model with confounders as covariates.

For all the PS methods, the wedge resection increases the probability of death and local recurrence. The type of surgery has a borderline effect on the non-local recurrence for all the methods analysed. The size of confidence intervals reflects the different sample size of the three endpoints: the most reliable estimations are the non-local recurrence ones.

4.4.3. Results on weighted all-cause hazard ratio

The previous analysis showed a strongly different effects of the two types of surgery on the three endpoints. Hence, a weighted all-cause hazard ratio is performed in order to estimate the surgical effect on a composite endpoint where the three endpoints have different weights. In particular, as suggested by the HERCOLES study clinicians, the local recurrence has a greater weight in terms of severity; therefore, the third scenario is the one with higher clinical interest (table 10).

Scenario	Weights
1. Standard approach	w_1 =1; w_2=1; w_3=1
2. Death is considered the worst event and the same relevance is given to the other endpoints	w_1 =0.5; w_2=0.5; w_3=1
3. Death is considered the worst event but also more relevance is given to local recurrence with respect to the other one	w_1 =0.8; w_2=0.5; w_3=1

Table 10. Description of the scenarios. Abbreviations: w1=weight for local recurrence; w2=weight for other recurrences; w3=weight for death

Among the PS based methods, for this part of the analysis, only the IPW and PS matching are compared because it is known that they are the two methods that minimize the bias in the estimation of the marginal hazard ratios [4].

Observing the weighted all-cause hazard ratio and its confidence intervals, provided by bootstrapping techniques (figure 24), the standard approach, both for IPW and PS matching, does not show significant difference between the two surgical techniques (scenario a). In the scenario b, there is a slight increase of the hazard ratio values but the confidence interval still does not highlight any significant difference between the two treatments.

The last scenario (scenario c) seems to suggest that the wedge resection increases the risk of the occurrence of one of the three events, in particular for the PS matching method.

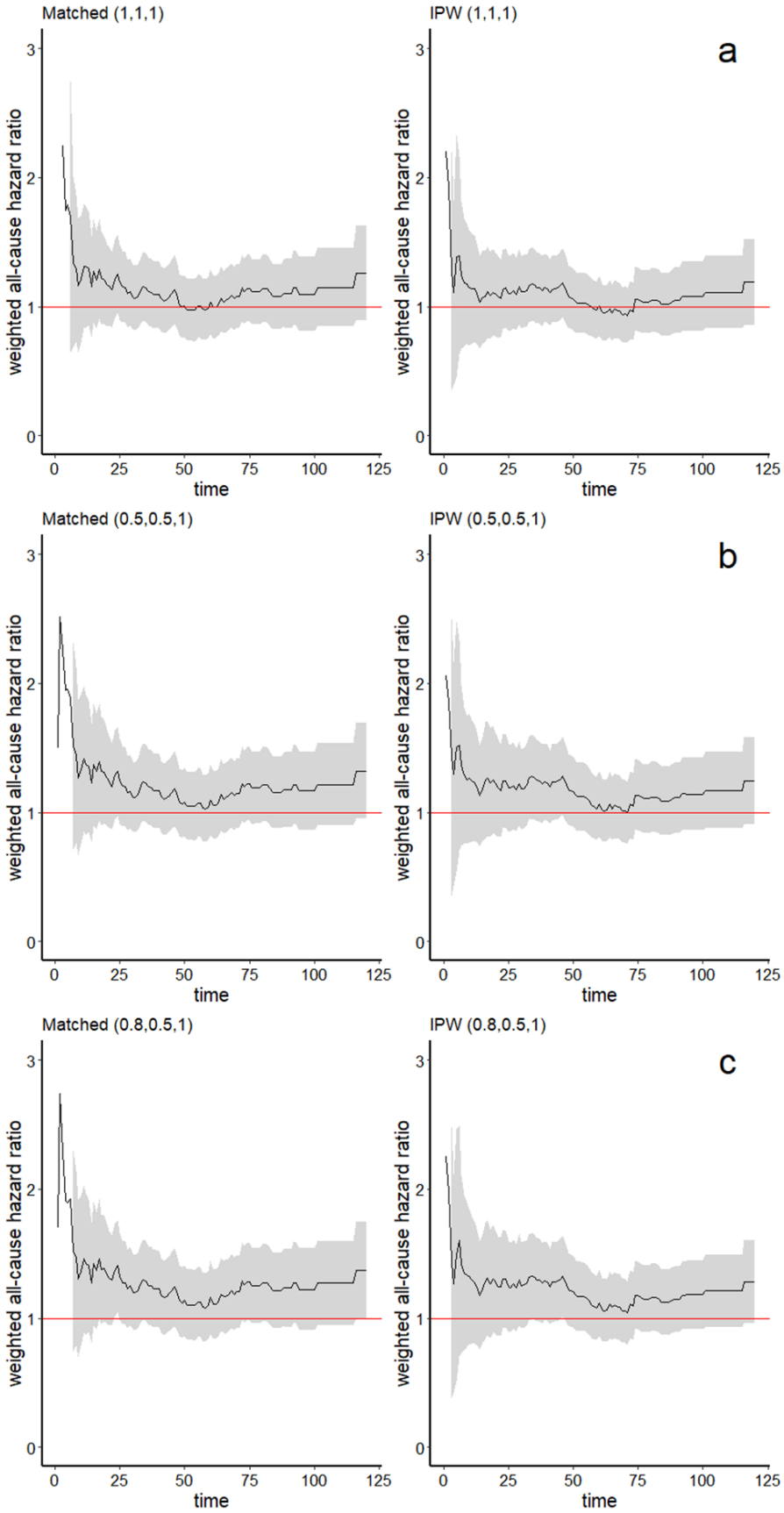


Figure 24. The weighted all-cause hazard ratio for Inverse probability weighting and propensity score matching in the three scenarios (a, b and c)

5. Discussion

Over the last few years, propensity score methods have been increasingly used in observational studies in medical literature to reduce confounding in estimating the treatment effect. Despite a wide collection of papers on propensity score analysis have been published, few of them are focused on the analysis of time to event outcomes [23].

The aim of this work was to add some novel considerations about the performance of different PS methods to estimate marginal hazard ratios and this was done considering a standard single endpoint, but also a composite endpoint. Another peculiar aspect of this work is the interest in dealing with composite endpoints. From a methodological point of view, the analysis of composite or single endpoints is not different, when the effect of the composite endpoint is based on the first event occurred irrespective of the type. However, sometimes it is of interest to assign a different clinical relevance to each single endpoint and thus a weighted effect measure, such as the weighted-all-cause-hazard-ratio [9], can be used to quantify the impact of treatment on the survival outcome. While there is some literature about the estimation of marginal hazard ratios for standard endpoints, the performance of PS methods to estimate causal effects on weighted endpoints is currently a topic still unexplored. Simulations studies were implemented to evaluate the behaviour of the various approaches we considered.

Results of the simulations on unweighted endpoints, confirm existing evidences about the effectiveness of IPW and matching on PS in reducing bias of the marginal treatment effect, while stratification and covariate adjustment using the PS result in estimation poorer performance [4][24][23]. More specifically, results here presented point out the primacy of the IPW over the other methods in terms of both precision (figures 6,9) and accuracy (figures 7,10) and thus a slightly worse performance of matching on the PS than IPW. This is not in agreement with all studies already present in the literature, e.g. [4]. A reason could be that the average number of matched pairs in this study (81.6%) is lower than in [4] (94.3%), causing a loss of both precision and accuracy of the estimates. A second limitation is the inclusion of only a single matching algorithm: greedy nearest-neighbour matching on the logit of the PS using calipers defined by the variance of the logit of the PS. This approach was used in this work because it has been found to perform well compared with other commonly used alternatives [17], but also other algorithms are available [16]. The higher biases observed for covariate adjustment and stratification using PS are also explainable. In fact, it

was demonstrated that these methods can estimate only conditional hazard ratios and not a marginal effect like matching or IPW [4].

The simulations on weighted all-cause hazard ratio present high computational complexity, therefore the two PS methods (stratification and covariate adjustment) were a priori excluded based on the biased results obtained on the unweighted context and only matching on PS and IPW were compared. The IPW results again the most precise (figures 11,14) and accurate (figures 12,15) method. Concerning PS matching, it has to be noted that the average number of matched pairs (70.3%) is lower than what obtained in the unweighted simulations. This could be, again, the main reason for lower efficiency associated with this method, compared to IPW. In all scenarios here considered, the assumption of equal cause-specific baseline hazards, which theoretically is required to guarantee that the non-parametric estimator of the all-cause-weighted-hazard-ratio of Ozga et al. [9], is violated. This was done because in practice (as in the motivating study here analysed), this assumption is often not tenable. Ozga et al. [9] state that the performance of the non-parametric estimator of the weighted all-cause hazard ratio remains good even when the assumption is not valid. Based on my results, I confirm this property of robustness of the weighted all-cause hazard ratio estimator also for the estimation of marginal treatment effects using PS matching or IPW.

In both simulation studies (dealing with unweighted and weighted endpoints) a limited number of scenarios was considered. Thus, the performance of the estimators of the marginal treatment effect could be different if tested in further situations and below different assumptions about the distribution of confounders and their relationship to treatment selection and outcomes. The setting of the parameters in this work was inspired by an application on a clinical observational study where the interest was to compare the effect of two surgical techniques to treat a particular liver tumor on the disease free survival (HERCOLES study).

The application of PS-based methods on the estimation of marginal hazard ratio for unweighted and weighted endpoints on the HERCOLES data is reported (Chapter 4). A common issue of propensity score based studies is the approach used to select confounders to be included in the PS model [25][26][27]. In this study, the confounding factors are identified using a logistic model with the least absolute shrinkage and selection operator (LASSO) variable selection method [22] that it was shown to be good in selecting true confounders and predictors of outcome [28]. Other automated variable selection methods have been criticized for the possibility of increasing risk of bias through over-adjustment on colliders or instrumental variables, but this issue still remains contentious in the current literature [25]. Another important issue that has to be taken into account when interpreting

the results of a PS-based analysis is the possible presence of unmeasured confounding. Some proposals have been recently presented in the literature to overcome this problem (e.g. the use of instrumental variables [29][30][31]) but further studies are needed to understand their relative performance in different contexts before they can be recommended. We assume that data regarding all the main relevant prognostic factor have been accurately measured in this study.

Initially, a standard (i.e. unweighted) composite endpoint was considered, namely the disease free survival (first event occurring among death, local and non-local recurrence). However, it was of interest to assign a different clinical relevance to each endpoint (death is considered obviously the worst event but also local recurrence is regarded as more severe than non-local one). Moreover, the competing risk analysis showed a strongly different effect of the two types of surgery on the three cause-specific events (figure 23). Hence, the non-parametric method of Ozga et al. [9] was adopted in order to estimate the surgical effect in terms of a weighted all-cause hazard ratio where the three endpoints have different weights. The choice of the weights for the different endpoints represents a controversial aspect for this method because it is somewhat arbitrary. A guidance for the choice of the relevance weights is proposed by Ozga et al. [9], but it is not definitive. The most important recommendation is that this process should be done in agreement with the study's clinicians, as it is done in this work.

Regarding the possible developments of this work, a good starting point could be the introduction of competing risk in this context, since actually there are very few guidelines on how to use propensity score based methods with competing risk data [32]. Another interesting point might be the comprehension of what happens when the assumption of independent censoring is violated: neither of the PS methods is suited to solve the problem of censored data when the censorship depends on the outcome [33]. In that case, an approach based on Inverse Probability of Censoring Weighting (IPCW) could be an option, provided that it is possible to accurately model the probability of censoring in time [34].

A final perspective could be to analyse the performance of the PS-based methods in estimating other marginal measures of effect (i.e. restricted mean survival time, time-fixed survival probability and survival quantile). With respect to the hazard ratio these measure have the advantage to be more easily interpretable from a clinical point of view [5] even though they are generally harder to estimate properly.

References

[1] P. R. Rosenbaum, “Propensity Score,” in *Encyclopedia of Biostatistics*, Chichester, UK: John Wiley & Sons, Ltd, 2005.

[2] E. Marubini and M. G. Valsecchi, *Analysing Survival Data from Clinical Trials and Observational Studies*. London, United Kingdom, 2004.

[3] M. A. Hernán and J. M. Robins, “Causal Inference: What If (2020),” 2020.

[4] P. C. Austin, “The performance of different propensity score methods for estimating marginal hazard ratios,” *Stat. Med.*, vol. 32, no. 16, pp. 2837–2849, 2013.

[5] H. Mao, L. Li, W. Yang, and Y. Shen, “On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference,” *Stat. Med.*, vol. 37, no. 26, pp. 3745–3763, Nov. 2018.

[6] P. C. Austin, “Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis,” *Stat. Med.*, vol. 35, no. 30, pp. 5642–5655, 2016.

[7] G. Rauch, K. Kunzmann, M. Kieser, K. Wegscheider, J. König, and C. Eulenburg, “A weighted combined effect measure for the analysis of a composite time-to-first-event endpoint with components of different clinical relevance,” *Stat. Med.*, vol. 37, no. 5, pp. 749–767, Feb. 2018.

[8] “Home | HERCOLESGroup.” [Online]. Available: https://www.hercolesgroup.eu/. [Accessed: 18-Oct-2019].

[9] A.-K. Ozga and G. Rauch, “Introducing a new estimator and test for the weighted all-cause hazard ratio.,” *BMC Med. Res. Methodol.*, vol. 19, no. 1, p. 118, 2019.

[10] M. Kleinbaum, David G. Klein, *Survival Analysis. A self-A self-Learning text*. 2012.

[11] D. A. Njamen-Njomen and J. Ngatchou-Wandji, “Nelson-Aalen and Kaplan-Meier Estimators in Competing Risks,” *Appl. Math.*, vol. 05, no. 04, pp. 765–776, 2014.

[12] O. Aaalen and S. Johansen, “An Empirical Transition Matrix for Non-homogeneous Markov Chains Based on Censored Observations,” *Scand. J. Stat.*, vol. 5, no. 3, pp. 141–150, 1978.

[13] J. A. Freiman, T. C. Chalmers, H. Smith, and R. R. Kuebler, “The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial:

Survey of 71 Negative Trials," *N. Engl. J. Med.*, vol. 299, no. 13, pp. 690–694, Sep. 1978.

[14] P. R. Rosenbaum and D. B. Rubin, "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *Am. Stat.*, vol. 39, no. 1, p. 33, Feb. 1985.

[15] W. G. Cochran, D. B. Rubin, S. Sankhyā, T. Indian, a Series, and B. W. G. Cochean, "Indian Statistical Institute Controlling Bias in Observational Studies : A Review All use subject to JSTOR Terms and Conditions CONTROLLING BIAS IN OBSERVATIONAL A REVIEW1 STUDIES :," vol. 35, no. 4, pp. 417–446, 2014.

[16] P. C. Austin, "A comparison of 12 algorithms for matching on the propensity score," *Stat. Med.*, vol. 33, no. 6, pp. 1057–1069, 2014.

[17] P. C. Austin, "Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies," *Pharm. Stat.*, vol. 10, no. 2, pp. 150–161, Mar. 2011.

[18] P. Lavrakas, "Propensity Scores," *Encycl. Surv. Res. Methods*, vol. 162, no. 8, pp. 734–737, 2013.

[19] M. A. Hernán and J. M. Robins, "Estimating causal effects from epidemiological data," *Journal of Epidemiology and Community Health*, vol. 60, no. 7. BMJ Publishing Group, pp. 578–586, Jul-2006.

[20] R. Bender, T. Augustin, and M. Blettner, "Generating survival times to simulate Cox proportional hazards models," *Stat. Med.*, vol. 24, no. 11, pp. 1713–1723, 2005.

[21] L. Giannitrapani *et al.*, "The Changing Epidemiology of Hepatocellular Carcinoma : Experience of a Single Center," *Biomed Res. Int.*, vol. 2020, p. 5309307, 2020.

[22] S. M. Kim, Y. Kim, K. Jeong, H. Jeong, and J. Kim, "Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography," *Ultrasonography*, vol. 37, no. 1, pp. 36–42, 2018.

[23] E. Gayat, M. Resche-Rigon, J. Y. Mary, and R. Porcher, "Propensity score applied to survival data analysis through proportional hazards models: A Monte Carlo study," *Pharm. Stat.*, vol. 11, no. 3, pp. 222–229, 2012.

[24] S. Deb *et al.*, “A Review of Propensity-Score Methods and Their Use in Cardiovascular Research,” *Can. J. Cardiol.*, vol. 32, no. 2, pp. 259–265, 2016.

[25] Y. K. Loke and K. Mattishent, “Propensity score methods in real-world epidemiology: A practical guide for first-time users,” *Diabetes, Obes. Metab.*, vol. 22, no. S3, pp. 13–20, 2020.

[26] J. A. Reiffel, “Propensity Score Matching: The ‘Devil is in the Details’ Where More May Be Hidden than You Know,” *Am. J. Med.*, vol. 133, no. 2, pp. 178–181, 2020.

[27] E. L. Fu, R. H. H. Groenwold, C. Zoccali, K. J. Jager, M. Van Diepen, and F. W. Dekker, “Merits and caveats of propensity scores to adjust for confounding,” *Nephrol. Dial. Transplant.*, vol. 34, no. 10, pp. 1629–1635, 2019.

[28] S. M. Shortreed and A. Ertefaie, “Outcome-adaptive lasso: Variable selection for causal inference,” *Biometrics*, vol. 73, no. 4, pp. 1111–1122, 2017.

[29] F. Torres, J. Ríos, J. Saez-Peñataro, and C. Pontes, “Is Propensity Score Analysis a Valid Surrogate of Randomization for the Avoidance of Allocation Bias?,” *Semin. Liver Dis.*, vol. 37, no. 3, pp. 275–286, 2017.

[30] H. Laborde-Castérot, N. Agrinier, and N. Thilly, “Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: A systematic review,” *J. Clin. Epidemiol.*, vol. 68, no. 10, pp. 1232–1240, 2015.

[31] A. J. Streeter *et al.*, “Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review,” *J. Clin. Epidemiol.*, vol. 87, pp. 23–34, 2017.

[32] P. C. Austin and J. P. Fine, “Propensity-score matching with competing risks in survival analysis,” *Stat. Med.*, vol. 38, no. 5, pp. 751–777, Feb. 2019.

[33] B. B. L. Penning De Vries and R. H. H. Groenwold, “Cautionary note: Propensity score matching does not account for bias due to censoring,” *Nephrology Dialysis Transplantation*, vol. 33, no. 6. Oxford University Press, pp. 914–916, 01-Jun-2018.

[34] J. M. Robins and D. M. Finkelstein, “Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests,” *Biometrics*, vol. 56, no. 3, pp. 779–788, 2000.

Appendix

In this Appendix the main lines of R code are reported for the simulation studies.

Simulation on unweighted hazard ratio

```
library(survival)
library(arm)
library(MatchIt)
library(splines)
library(gtools)

Nsim=10000
for (j in 1:Nsim) {
 N=1000
 x1   <-  rbinom(N, 1, prob=0.70)
 alpha1 <- 0.53
 x2   <-  rpois(N, 1.3)
 alpha2 <- 0.38
 x3   <-  rnorm(N,75,10)
 alpha3 <- -0.01
 x4 <- rbinom(N,1,0.1)
 alpha4 <- 0.49
 x5 <- rbinom(N, 1, prob=0.51)
 alpha5 <- -0.08
 x6   <-  rnorm(N,1.2,0.24)
 alpha6 <- -1
 x7   <-  rpois(N, 180)
 alpha7 <- -0.01
 x8   <-  rnorm(N,4.6,2.9)
 alpha8 <- -0.23
 x9.2 <- rbinom(N, 1, prob=0.66)
 alpha9.2 <- -0.35
```

```
x9.3 <- ifelse(x9.2==1,0,rbinom(1, 1, prob=0.73))
alpha9.3 <- 0.02
x10 <- rbinom(N,1,0.02)
alpha10 <- 0.77
x11 <- rbinom(N,1,0.48)
alpha11 <- 0.59
alpha0 <- 3.2
p_wedge <- invlogit(alpha0+alpha1*x1+alpha2*x2+alpha3*x3+alpha4*x4+ alpha5*x5+
alpha6*x6+alpha7*x7+alpha8*x8+alpha9.2*x9.2+alpha9.3*x9.3+alpha10*x10+alpha11*x11)
CreateTreat <- function(N, var) {rbinom(N,1,var)}
treatment <- CreateTreat(1000,p_wedge)
table(treatment)
t.cens <- runif(N, min=360, max=500)
U<-runif(N, min=0, max=1)
k<-0.15
p<-1
beta1 <- 0.19
beta4 <- 0.16
beta9.2 <- 0.72
beta9.3 <- 1.49
beta10 <- 0.65
beta11 <- 0.39
beta2 <- 0.22
beta6 <- -1.78
beta7 <- -0.01
beta8 <- -0.02
beta3 <- -0.01
beta5 <- -0.13
HR <-  1.5
beta0 = 0.501

iter = 0
```

```
 repeat {
  iter = iter+1
  t.event0 <-  (-log(U))/(k*exp(beta1*x1+beta2*x2+beta3*x3+beta4*x4+beta5*x5+
   beta6*x6+beta7*x7+beta8*x8+beta9.2*x9.2+beta9.3*x9.3+beta10*x10+beta11*x11))^(1/p)
  t.event1 <-  (-log(U))/(k*exp(beta0+beta1*x1+beta2*x2+beta3*x3+beta4*x4+beta5*x5+
   beta6*x6+beta7*x7+beta8*x8+beta9.2*x9.2+beta9.3*x9.3+beta10*x10+beta11*x11))^(1/p)
  mod <- coxph(Surv(c(t.event0,t.event1),rep(1,2*N))~c(rep(0,N),rep(1,N)))
  if (round(summary(mod)$coef[1,1],3) == round(log(HR),3) | iter==1000) break
  if (summary(mod)$coef[1,1] < log(HR)) {beta0 <- beta0 + 0.001}
  if (summary(mod)$coef[1,1] > log(HR)) {beta0 <- beta0 - 0.001}
 }
 if (iter ==1000) {salti = salti+1
 next
 }
 # observed time and event indicators
 t.event <- ifelse(treatment==1, t.event1, t.event0)
 T.event<-pmin(t.event,t.cens)
 Event<-ifelse(t.event<t.cens,1,0)
 data_sim <- data.frame(T.event,Event,treatment, x1,x2,x3,x4,x5,x6,x7,x8,x9.2,x9.3,x10,x11)

 #calculation of weights
 mod1 <- glm(treatment ~ x1+x2+x3+x4+x5+x6+x7+x8+x9.2+x9.3+x10+x11, data=data_sim,
family="binomial")
 data_sim$ps_value <- predict(mod1, type="response")
 data_sim$ipw <- ifelse(data_sim$treatment==1, 1/data_sim$ps_value, 1/(1-data_sim$ps_value))
 mod2 <- glm(treatment ~ 1, data=data_sim, family="binomial")
 data_sim$p_treat <- predict(mod2, type="response")
 data_sim$ipw_st <- ifelse(data_sim$treatment==1,data_sim$p_treat/(data_sim$ps_value),(1-
data_sim$p_treat)/(1- data_sim$ps_value))

 #IPW
 mod4 <- coxph(Surv(T.event,Event)~ treatment, weights=ipw, data=data_sim, robust=T)
```

```
  #Matching
  match <- matchit(treatment ~ x1+x2+x3+x4+x5+x6+x7+x8+x9.2+x9.3+x10+x11, data=data_sim,
                   method = "nearest", caliper=0.2)
  data_complete_match <- match.data(match)
  mod6 <- coxph(Surv(T.event,Event)~ treatment,data=data_complete_match, robust=T)

  #PS as covariate with spline transformation
  mod8 <- coxph(Surv(T.event,Event)~ treatment + ns(ps_value,df=4), data=data_sim, robust=T)

  #Stratification on PS quintiles
  data_sim$ps_quintile <- quantcut(data_sim$ps_value, q=5, format="d")
  mod9 <- coxph(Surv(T.event,Event)~ treatment + as.factor(ps_quintile), data=data_sim,
robust=T)

  print(j)
}
```

Simulation on weighted all-cause hazard ratio

```
library(survival)
library(arm)
library(MatchIt)
library(splines)
library(gtools)
library(dplyr)
library(tidyr)
library(dynpred)



###################################
#IPW
Nsim=1000
for (j in 1:Nsim) {
 N=1000
 x1   <-  rbinom(N, 1, prob=0.70)
 alpha1 <- 0.53
 x2   <-  rpois(N, 1.3)
 alpha2 <- 0.38
 x3   <-  rnorm(N,75,10)
 alpha3 <- -0.01
 alpha0 <- 0.1

 p_wedge <- invlogit(alpha0+alpha1*x1+alpha2*x2+alpha3*x3)
 CreateTreat <- function(N, var) {rbinom(N,1,var)}
 treatment <- CreateTreat(1000,p_wedge)
 table(treatment)

 k<-0.5
 p<- 1
```

```
l<-0.7
q<- 2
beta1 <- 0.19
beta2 <- 0.22
beta3 <- -0.01
beta0.1=1
beta0.2=1
w1=1
w2=0.5
cumHR=2

iter = 0
repeat {
 iter = iter+1
 U<-runif(N, min=0, max=1)
 t.event1A<- (-log(U))/(k*exp(beta1*x1+beta2*x2+beta3*x3))^(1/p)
 U<-runif(N, min=0, max=1)
 t.event2A<- (-log(U))/(l*exp(beta1*x1+beta2*x2+beta3*x3))^(1/q)
 U<-runif(N, min=0, max=1)
 t.event1B <- (-log(U))/(k*exp(beta0.1+beta1*x1+beta2*x2+beta3*x3))^(1/p)
 U<-runif(N, min=0, max=1)
 t.event2B <- (-log(U))/(l*exp(beta0.2+beta1*x1+beta2*x2+beta3*x3))^(1/q)

 status_A <- ifelse(t.event1A<t.event2A,1,2)
 status_B <- ifelse(t.event1B<t.event2B,1,2)
 t.eventA <- pmin(t.event1A,t.event2A)
 t.eventB <- pmin(t.event1B,t.event2B)

 mod1A <- summary(survfit(Surv(t.eventA,status_A==1)~1))
 mod2A <- summary(survfit(Surv(t.eventA,status_A==2)~1))
 mod1B <- summary(survfit(Surv(t.eventB,status_B==1)~1))
 mod2B <- summary(survfit(Surv(t.eventB,status_B==2)~1))
```

```
  h1A <- mod1A$n.event / mod1A$n.risk
  h2A <- mod2A$n.event / mod2A$n.risk
  h1B <- mod1B$n.event / mod1B$n.risk
  h2B <- mod2B$n.event / mod2B$n.risk
  cum1A <- sum(h1A[mod1A$time<1])
  cum2A <- sum(h2A[mod2A$time<1])
  cum1B <- sum(h1B[mod1B$time<1])
  cum2B <- sum(h2B[mod2B$time<1])
  cumHRcond <- (w1*cum1B + w2*cum2B)/(w1*cum1A + w2*cum2A)
  if (round(cumHRcond,3) == round(cumHR,3)) break
  if (cumHRcond < cumHR) {beta0.2 <- beta0.2 + 0.001}
  if (cumHRcond > cumHR) {beta0.2 <- beta0.2 - 0.001}
 }

 #observed time and event indicators
 t.cens <- runif(N, min=0.5, max=2)
 t.event <- ifelse(treatment==1, t.eventB, t.eventA)
 T.event<-pmin(t.event,t.cens)
 event<- ifelse(treatment==1, status_B, status_A)
 Event<-ifelse(t.event<t.cens,event,0)
 data_sim <- data.frame(T.event,Event,treatment, x1,x2,x3)

 #calculation of weights
 mod1 <- glm(treatment ~ x1+x2+x3, data=data_sim, family="binomial")
 data_sim$ps_value <- predict(mod1, type="response")
 data_sim$ipw <- ifelse(data_sim$treatment==1, 1/data_sim$ps_value, 1/(1-data_sim$ps_value))

 mod2 <- glm(treatment ~ 1, data=data_sim, family="binomial")
 data_sim$p_treat <- predict(mod2, type="response")
 data_sim$ipw_st <- ifelse(data_sim$treatment==1,data_sim$p_treat/(data_sim$ps_value),(1-
data_sim$p_treat)/(1- data_sim$ps_value))
 ##anatomic resection
```

```
 anatomic <- data_sim[which(data_sim$treatment==0),]
 #Event 1
 mod <- summary(survfit(Surv(T.event,Event==1)~1,data=anatomic,weights=ipw))
 hazard <- mod$n.event / mod$n.risk
 IPW1A <- sum(hazard[mod$time<1])
 #Event 2
 mod <- summary(survfit(Surv(T.event,Event==2)~1,data=anatomic,weights=ipw))
 hazard <- mod$n.event / mod$n.risk
 IPW2A <- sum(hazard[mod$time<1])

##wedge resection
 wedge <- data_sim[which(data_sim$treatment==1),]
 #Event 1
 mod <- summary(survfit(Surv(T.event,Event==1)~1,data=wedge,weights=ipw))
 hazard <- mod$n.event / mod$n.risk
 IPW1B <- sum(hazard[mod$time<1])
 #Event 2
 mod <- summary(survfit(Surv(T.event,Event==2)~1,data=wedge,weights=ipw))
 hazard <- mod$n.event / mod$n.risk
 IPW2B <- sum(hazard[mod$time<1])
 ##calculation of weighted all-cause HR
 wHR_ipw <- (IPW1B*w1+IPW2B*w2)/(IPW1A*w1+IPW2A*w2)
 HR <- c(HR,wHR_ipw)
 n.cens <- c(n.cens, table(data_sim$Event)[1])
 n.1event <- c(n.1event, table(data_sim$Event)[2])
 n.2event <- c(n.2event, table(data_sim$Event)[3])

 print(j)
}
##########################################
#MATCHING
Nsim=1000
```

```
for (j in 1:Nsim) {
 N=1000
 x1   <-  rbinom(N, 1, prob=0.70)
 alpha1 <- 0.53
 x2   <-  rpois(N, 1.3)
 alpha2 <- 0.38
 x3   <-  rnorm(N,75,10)
 alpha3 <- -0.01
 alpha0 <- 0.1

 p_wedge <- invlogit(alpha0+alpha1*x1+alpha2*x2+alpha3*x3)
 CreateTreat <- function(N, var) {rbinom(N,1,var)}
 treatment <- CreateTreat(1000,p_wedge)
 table(treatment)

 k<-0.5
 p<- 1
 l<-0.7
 q<- 2
 beta1 <- 0.19
 beta2 <- 0.22
 beta3 <- -0.01
 beta0.1=1
 beta0.2=1
 w1=1
 w2=0.5
 cumHR=2
 iter = 0

 repeat {
  iter = iter+1
  U<-runif(N, min=0, max=1)
```

```
  t.event1A<- (-log(U))/(k*exp(beta1*x1+beta2*x2+beta3*x3))^(1/p)
  U<-runif(N, min=0, max=1)
  t.event2A<- (-log(U))/(l*exp(beta1*x1+beta2*x2+beta3*x3))^(1/q)
  U<-runif(N, min=0, max=1)
  t.event1B <- (-log(U))/(k*exp(beta0.1+beta1*x1+beta2*x2+beta3*x3))^(1/p)
  U<-runif(N, min=0, max=1)
  t.event2B <- (-log(U))/(l*exp(beta0.2+beta1*x1+beta2*x2+beta3*x3))^(1/q)

  status_A <- ifelse(t.event1A<t.event2A,1,2)
  status_B <- ifelse(t.event1B<t.event2B,1,2)
  t.eventA <- pmin(t.event1A,t.event2A)
  t.eventB <- pmin(t.event1B,t.event2B)

  mod1A <- summary(survfit(Surv(t.eventA,status_A==1)~1))
  mod2A <- summary(survfit(Surv(t.eventA,status_A==2)~1))
  mod1B <- summary(survfit(Surv(t.eventB,status_B==1)~1))
  mod2B <- summary(survfit(Surv(t.eventB,status_B==2)~1))
  h1A <- mod1A$n.event / mod1A$n.risk
  h2A <- mod2A$n.event / mod2A$n.risk
  h1B <- mod1B$n.event / mod1B$n.risk
  h2B <- mod2B$n.event / mod2B$n.risk
  cum1A <- sum(h1A[mod1A$time<1])
  cum2A <- sum(h2A[mod2A$time<1])
  cum1B <- sum(h1B[mod1B$time<1])
  cum2B <- sum(h2B[mod2B$time<1])
  cumHRcond <- (w1*cum1B + w2*cum2B)/(w1*cum1A + w2*cum2A)
  if (round(cumHRcond,3) == round(cumHR,3)) break
  if (cumHRcond < cumHR) {beta0.2 <- beta0.2 + 0.001}
  if (cumHRcond > cumHR) {beta0.2 <- beta0.2 - 0.001}
}
```

```
#observed time and event indicators
t.cens <- runif(N, min=0.5, max=2)
t.event <- ifelse(treatment==1, t.eventB, t.eventA)
T.event<-pmin(t.event,t.cens)
event<- ifelse(treatment==1, status_B, status_A)
Event<-ifelse(t.event<t.cens,event,0)
data_sim <- data.frame(T.event,Event,treatment, x1,x2,x3)

match <- matchit(treatment ~ x1+x2+x3, data=data_sim, method = "nearest",caliper=0.2)
data_complete_match <- match.data(match)
##anatomic resection
anatomic <- data_complete_match[which(data_complete_match$treatment==0),]
#Event 1
mod <- summary(survfit(Surv(T.event,Event==1)~1,data=anatomic))
hazard <- mod$n.event / mod$n.risk
M1A <- sum(hazard[mod$time<1])
#Event 2
mod <- summary(survfit(Surv(T.event,Event==2)~1,data=anatomic))
hazard <- mod$n.event / mod$n.risk
M2A <- sum(hazard[mod$time<1])
##wedge resection
wedge <- data_complete_match[which(data_complete_match$treatment==1),]
#Event 1
mod <- summary(survfit(Surv(T.event,Event==1)~1,data=wedge))
hazard <- mod$n.event / mod$n.risk
M1B <- sum(hazard[mod$time<1])
#Event 2
mod <- summary(survfit(Surv(T.event,Event==2)~1,data=wedge))
hazard <- mod$n.event / mod$n.risk
M2B <- sum(hazard[mod$time<1])
##calculation of the all-cause HR
wHR_match <- (M1B*w1+M2B*w2)/(M1A*w1+M2A*w2)
```

```
  HR <- c(HR,wHR_match)
  n.cens <- c(n.cens, table(data_sim$Event)[1])
  n.1event <- c(n.1event, table(data_sim$Event)[2])
  n.2event <- c(n.2event, table(data_sim$Event)[3])
  dim.match <-c(dim.match, dim(data_complete_match)[1])

  print(j)
}
```