



Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Data Science

**Propensity Score Methods to Estimate the
Average Causal Effect with Multiple Treatments:
Application to the Study of Surgical Resection
Techniques in Liver Carcinoma
and Simulation Study**

Relatore: Dott. Davide Paolo Bernasconi

Correlatrice: Dott.ssa Giulia Capitoli

Tesi di Laurea Magistrale di:

Anna Nava

827719

Anno Accademico 2021–2022

Abstract

At the top of the global health challenges, the hepatocellular carcinoma (HCC) - which is the most common type of liver cancer - has an alarming mortality rate. So far, the first-line therapy is surgery, which has nevertheless unsatisfying results in the long term period, during which the majority of the patients has a relapse.

To improve the efficacy of surgical interventions, the project HE.RC.O.LE.S. was set: it is a nation-wide collaborative network of Italian centers that shares the data of patients operated after a diagnosis of HCC. Sourcing from this Register of data, we realised a causal inference study which compares the effects of three resection techniques - the anatomic, the semi-anatomic and the wedge resection. In particular, our objective is determining which of them is the best to operate patients affected by large HCC (i.e. with at least one nodule of diameter >5 cm). Unfortunately, it emerged that, even within this relatively homogeneous population, it's not possible to identify a treatment generically better than the others: the optimal resection technique can be found on the basis of patients' characteristics, besides the tumour dimensions.

In the second part of the thesis, we set up a simulation study to test the performances of the propensity score methods previously used.

In the scientific literature, there is paucity of papers about how to estimate causal effects from observational studies that concern more than two treatments, so it was of particular interest to investigate if standard binary-exposure propensity score models still work accurately and efficiently with three categories of therapy.

The methods compared are the adjusted Cox model, the propensity score matching model, the inverse probability weighting using the propensity score model and the generalised propensity score cumulative distribution function model. The latter is a method built specifically for multi-treatments settings.

In our generated cases, the last model, which is also the most complicated, was outdone in terms of performances by the other more classical propensity score methods. We concluded that adapting methods usually utilised with binary treatments already brings to valid results in three exposures settings, similar to the one we considered. Further studies could analyse what would happen with more than three therapies groups.

Contents

1	Introduction	1
1.1	Objectives	2
2	Methods	3
2.1	Causal Inference in clinical researches	3
2.2	Survival Analysis: basic functions	6
2.3	Parametric models	7
2.4	Non-parametric models	9
2.5	Survival analysis main issues	10
2.6	Cox model	11
2.6.1	Parameters estimation	11
2.7	Log-rank test	12
2.8	Estimations of treatment effects	13
2.9	Propensity score methods	15
2.9.1	Matching	16
2.9.2	Inverse probability weighting	17
2.9.3	Generalised propensity score cumulative distribution function	18
2.10	Models for the estimation of propensity score	20
2.10.1	Multinomial logistic regression	20
2.10.2	Generalised boosted regression	22
3	Case of study	23
3.1	Clinical context	23
3.2	Dataset	25
3.3	Multiple imputation	33
3.4	Applied methods	35
3.5	Results	43
3.6	Discussion	50
4	Simulation study	53
4.1	Monte Carlo method	53
4.2	Study design	53
4.3	Results	55
4.4	Discussion	65
5	Conclusion	67

*For thousands more years
the mighty ships tore across the empty wastes of space
and finally dived screaming on to the first planet they came across
- which happened to be the Earth -
where due to a terrible miscalculation of scale
the entire battle fleet was accidentally swallowed by a small dog.
Those who study the complex interplay of cause and effect in the history
of the Universe say that this sort of thing is going on all the time,
but that we are powerless to prevent it.
“It’s just life,” they say.*

— The Hitchhiker’s Guide to the Galaxy, Douglas Adams

1. Introduction

The hepatocellular carcinoma (HCC) is the most widely spread liver primary tumour and one of the most common causes of cancer death worldwide. In the attempt of defeating this disease, the first-line therapeutic option is surgery. However, this is feasible only in the 20% or 30% of patients, due to the poor hepatic reserve function caused by the neoplasia. Moreover, the recurrence rate after the surgical intervention remains high: 50% during the first 3 years and $>70\%$ during the first 5 years. In this scenario, it is crucial to investigate the long term effects of the therapies adopted, especially verifying if the choice of the resection technique is relevant to avoid future complications. For this reason, in this thesis, we will apply survival analysis methods to compare the outcomes of three types of surgery: the anatomic, which cuts a relatively large area with more chances of removing the whole cancer; the wedge, which is the most conservative one, trying to cut only the tumour area; and the semi-anatomic – the middle way. This work, which concerns an observational multi-centric retrospective cohort study, relies on a dataset of 613 patients affected by large HCC, i.e. a tumour with at least one nodule of diameter >5 cm. The data belongs to an Italian Register, fruit of the work of the study group of the HE.RC.O.LE.S. project: their objective is to collect data about surgical treatments for HCC from various clinics all over the country, with the intent of making it available to set up deep and comprehensive analyses. This register contains the participants' clinical details, essential anagraphic data - that has obviously been anonymized -, date, hospital, type of surgical operations and time-to-event values.

Since our analysis consists in an observational study, we have to face the problem that the treatment assignment is not randomized, which implies that the considered patients' groups are not homogeneous, so there are numerous confounding factors. To limit this problem, propensity score methods are utilised: they balance the distribution of observed baseline covariates between treated and untreated subjects, to mimic a randomized controlled trial.

Another issue that arises is the fact that in the scientific literature, there are many and exhaustive papers about estimations of the causal effects in observational studies with two treatments, but not with multiple treatments. For this reason, there is no clear guidance on which causal inference technique performs best if you want to compare more than two therapies. Since in this case we are considering *three* techniques of resection, in addition to the objective of comparing them in the case of large HCC and analysing their long term effects, we pursue the goal of testing some of the propensity score methods proposed specifically to deal with multiple treatments.

In particular, we will implement on R the following models: the univariate Cox model, the covariate adjusted Cox model, the inverse probability of treatment weighting (IPW) using the propensity score and the generalised propensity score cumulative distribution function (GPS-CDF) model. All of them estimate the ATE considering all the resection

types at the same time.

Then, we will perform two separate analyses: the first one comparing subjects treated with anatomical versus semi-anatomical resections and the second one comparing subjects treated with anatomical versus wedge resections. In this way, the multi-treatment analyses are rethought as a composition of two binary-treatment studies and the Average Treatment Effect in the Treated (ATT) is estimated for both comparisons separately, by standard propensity score matching models.

In IPW, the propensity score will be estimated in two ways: firstly, with a multinomial regression and then with a generalised boosted regression.

At the end, we will compare six models: the unadjusted Cox model, the adjusted Cox model, the IPW multinomial model, the IPW boosted model, the GPS-CDF model and the standard matching model.

The thesis is composed by six chapters, whose content is here illustrated.

This is the first introductory chapter, with a summary of the main topics of the work.

Then, the second chapter begins with a presentation of the basic knowledge about causal inference and survival analysis. It continues with an in-depth description of the propensity score methods utilised.

The following section named "Case of study" gives an overview of the clinical contest and shows the descriptive analyses of the dataset and the illustration of data cleaning phase. At this point, the models explained in the previous chapter can be finally applied to the large HCC dataset: the results are reported and commented.

In the next chapter, we conduct a simulation study, generating 1000 datasets by Monte Carlo method, each composed by 1000 subjects. They are analysed with the same propensity score methods of the chapter 3: the aim is to test the efficiency and the bias of the methods with respect to the true marginal effect.

The thesis ends with a conclusion chapter that summarizes the main findings and makes reflections about them.

Depicted the clinical contest and given an overview of the work content, let's now illustrate the thesis goals.

1.1 Objectives

This thesis is pursuing two main objectives.

It is born with the goal of analysing the long term effects of different resection techniques utilised to cure large HCC. One of them is expected to work generally better than the others: to investigate this hypothesis, we apply propensity score methods to our data. Since the surgery categories compared are three and so, we are not considering the standard binary exposure setting, another research task emerges: we aim to investigate the propensity score methods performances in the study of three treatments. In particular, we wonder if classical approaches are still effective and accurate or if it is necessary to utilise ad-hoc models.

2. Methods

2.1 Causal Inference in clinical researches

Before entering into the merits of the study, it's necessary to dedicate some time to the statistics knowledge at the basis of survival analysis. This entire section offers an overview of the principal concepts required to understand the rest of the paper.

The field we are examining concerns biomedical research, whose aim is to establish the relationship between a characteristic or a treatment and a disease, in particular looking for **cause-effect** relationships. In fact, while many correlations between data exist, most of them are just spurious: *correlation does not imply causation*.

At this point, a crucial question arises: how to prove the existence of causality? No simple answers exist. In 1915, Sir Austin Bradford Hill suggested nine aspects of association to consider for the purpose of deciding if the association can be interpreted as a cause-effect relationship [1]. Today, they are known as Bradford Hill Criteria and they are often considered as reference points in causal inference studies in epidemiology, even if they've been obviously revisited and modernized during the years.

The nine "viewpoints", as Hill called them, are the following:

1. **Strength**: the stronger it is an association, the more likely it is to be causal. Nowadays, this strength is judged by its statistical significance.
2. **Consistency**: to be consistent, an association should be tested by multiple studies with variety of populations, locations and circumstances. Indeed, by repetitions of observations, the researcher should verify that the statistical significance hasn't been found by chance.
3. **Specificity**: if the exposure causes only one disease, it is an argument in favour to the causation. Anyway, this point is considered obsolete.
4. **Temporality**: the demonstration that the exposure precedes the beginning of the disease helps to prove that there is a causal relationship between them. Even if this aspect can appear obvious, there is often the risk of confusing the cause with the effect and vice versa.
5. **Biological gradient**: here there is another outdated criteria. Hill states that the presence of a linear dose-response curve suggests causality, but with modern knowledge it's clear that this is an over-simplification.
6. **Plausibility**: a biological possible explanation of the association is a support for the thesis of causation. Anyway, Hill himself recognized the limits of this point: plausibility is strictly connected to the knowledge of the day, which is continuously in evolution.

7. **Coherence:** really similar to the previous criteria, it refers to the fact that associations should be coherent and shouldn't seriously conflict with the general available knowledge.
8. **Experiment:** Hill hypothesizes that experimental evidences may be the strongest support for causal inference. Although the importance of experiments is undeniable, they are often impossible to realize, as we will discuss later.
9. **Analogy:** given an association which has been proved to be causality, analogous ones are more likely to be cause-effect relationships. Even this point is pretty weak, since it's very vague.

Anyway, not all Bradford Hill Criteria are still actual; an interesting discussion on how they can be recontextualized and applied nowadays is done in "Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology" [2]. However, there is even who thinks that they should be *completely* abandoned. For instance, Rothman and Greenland harshly criticize them, stating that no absolute criteria exists for verifying the validity of scientific evidences; the unique possibility is to assess the validity of an entire study [3]. The quality of a scientific research can indeed be estimated by the total errors committed in the whole work. In fact, since it is impossible to completely avoid the errors, a scientist should aim for minimizing them.

Despite of this conflicting subject, which involve metaphysical and philosophical issues, the scientific community recognizes a shared gold standard in causal inference, which is randomised controlled trial (RCT) - confirming Hill's ninth point about experiments. It has in fact various characteristics that are advantageous for the researcher.

First of all, in experimental studies, the investigators directly assign the exposure, which is in this way under their control. Moreover, randomisation has great benefits: it forms the basis of statistical tests, the selection bias is eliminated, the covariates distributions are homogeneous between the treated subjects' and the untreated subjects' group and, consequently, the presence of confounding variables is strongly limited [4].

Unfortunately, there is a "but": in medical researches, it's often not possible realising a RCT, due to ethical or practical reasons. This is why observational studies are so important: like their name suggests, they are defined as studies where the researchers monitor the effect of an exposure *passively*, that is they don't decide and don't change who is or who isn't exposed. This aspect makes the analyses more difficult, yet not impossible.

Going deeper in the taxonomy of clinical researches 2.1, two main subgroups of observational studies can be defined: descriptive studies and analytical studies, distinguishable from the fact that the first ones don't have comparison groups. The analytical studies can be similarly divided into three categories [5]:

- **Cross-sectional studies:** at a certain point in time, the investigators look at the considered population and register the number of who resulted to be treated and the number of who presents a disease. Therefore, these studies concern *prevalence* and not *incidence*. The latter refers to the number of new cases of a specified event

developed during a particular time period.

Note that the principle of *temporality* can't be pursued, because causes and effects are considered all at the same time.

- **Case-control studies:** the observations go backwards, starting from the outcomes in search of the causes. In particular, the investigators consider a group of people who present the outcome - the cases - and a group of people without the outcome - the controls. Then, the researchers verify the prevalence of an exposure for both groups: if it is higher in the group of cases, the exposure is considered responsible of an increased risk of the outcome.

They are particularly suitable for outcomes rare or that take long time to develop. Their main flaws are instead the difficulty of choosing controls properly - and doing that badly can ruin the whole study -, and the risk of recall bias.

- **Cohort studies:** chosen a cohort, a starting point and an exposure, the investigators observe the study participants, looking forwards for the outcomes. In particular, the cohort is composed by a group of individuals who share one or more stabilized characteristics. Unlike case-control, these studies permit to compute incidence rates and relative risks, but they take very long time and implicate high costs and effort.

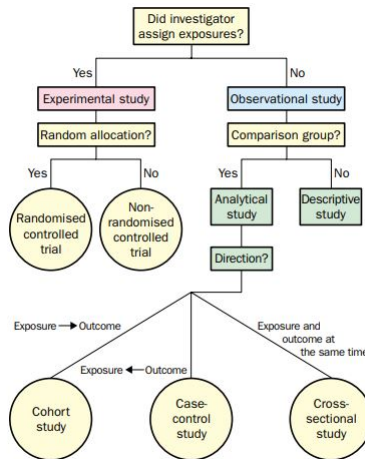


Figure 2.1: Taxonomy of clinical researches, from [5]

The case we consider belongs to the cohort studies, where the examined population is composed by patients who had large HCC and, consequently, were operated. The day of the surgical intervention is the starting point, while the exposures that are compared are the types of resection utilised, grouped in three categories. Since the outcomes observed are time-to-event, let's have an overview about **survival analysis**.

2.2 Survival Analysis: basic functions

Survival analysis is an analysis of data that involves times to an event of interest [6]. Firstly, a starting point is needed to be established: it defines the moment after which the patients are observed. The event, which is usually the patients' death or their relapse, determines the study end-point.

Every participant has their period of follow-up, whose conclusion can be motivated with three reasons:

- the patient had the event;
- the patient is *lost to follow-up*, due to external causes;
- the patient has been administrative censored: at the cut-off date of the analysis, they still didn't have the event.

Let's now define the functions that are essential in these analyses.

Definition 2.2.1. The **survival function** $S(t)$ is the probability of surviving after the time t . Said the random variable $T \geq 0$ the survival time,

$$S(t) = P(T > t) = \int_t^\infty f(u)du$$

$S(t)$ is monotone decreasing function, equal to one at time zero. The complementary of the survival function is the **incidence function**.

Definition 2.2.2. The **incidence function** $F(t)$ is the probability of failing before the time t , such that

$$F(t) = P(T \leq t) = 1 - \int_t^\infty f(u)du = \int_0^t f(u)du$$

Clearly, $F(t)$ is monotone increasing function, equal to zero at time zero. Starting from these concepts, it's possible to define one of the core function of the survival analysis.

Definition 2.2.3. The **hazard function** $\lambda(t)$ indicates the velocity of event development in event-free subjects:

$$\lambda(t) = \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$

The hazard function, differently from survival and incidence ones, isn't monotonic. Its trend is meaningful about the risk of death of the considered subjects. Its integral is called **cumulative hazard function**:

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

It is related to the survival function by the following equation:

$$S(t) = \exp -\Lambda(t) \quad (2.1)$$

To demonstrate that, let's first look at the definition of the hazard:

$$\lambda(t) = \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{dF(t)}{S(t)} = -\frac{d(1 - F(t))}{S(t)} = -\frac{dS(t)}{S(t)} = -d \log(S(t))$$

Then, integrate both sides of the equation above:

$$\log(S(t)) = \int_0^t \lambda(u) du$$

At the end, we've found out (2.1).

These functions are what survival analysis aims to estimate. In particular, there are three possible categories of methods to do so: **parametric models**, **non-parametric models** and **semi-parametric models**. While the first ones make assumptions of the survival time distribution, the others don't. Finally, the semi-parametric models are characterized by the fact that some of their regression parameters are known ([7],ch.3.1).

2.3 Parametric models

Given a dataset, the idea of these models is usually to make assumptions of the shape of the hazard function. They are in particular based on the estimations of some parameters, which can be derived from the data.

The distributions usually considered are ([6]; [8],ch.7):

- The **exponential distribution**, which is the simplest one.
There is only one parameter to estimate - h -, while the hazard and the survival function are assumed to be equal to:

$$\lambda(t) = h, S(t) = \exp(-ht)$$

Note that we are making a strong assumption defining the hazard as constant: it implies that the risk of death doesn't change with time.

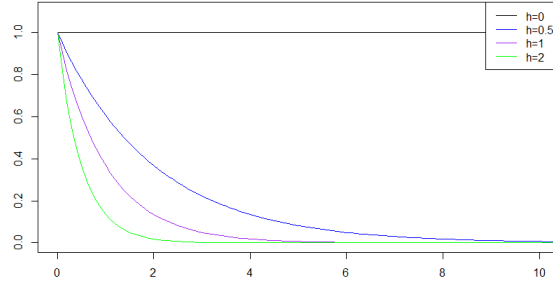


Figure 2.2: Survival functions along time fixed the hazard

- The **Weibull distribution**, for which a new parameter is introduced - p - and the function equations become:

$$\lambda(t) = hpt^{p-1}, S(t) = \exp(-ht^p)$$

So, the exponential distribution is a special case of this, with $p = 1$. The hazard now follows a monotonic trend, increasing or decreasing.

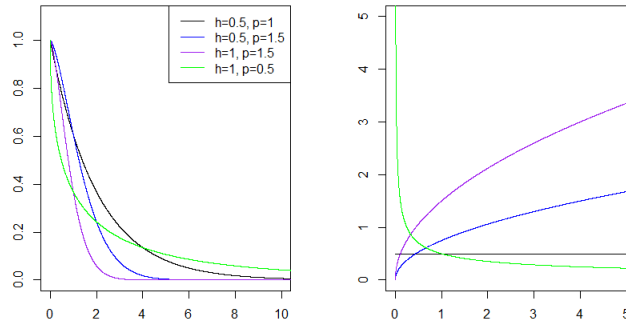


Figure 2.3: Survival and hazard functions along time, fixed the parameters

- The **log-logistic distribution**, that still uses two parameters and is based on the equations:

$$\lambda(t) = \frac{hpt^{p-1}}{1+ht^p}, S(t) = \frac{1}{1+ht^p}$$

The hazard can either be a monotonic function or unimodal.

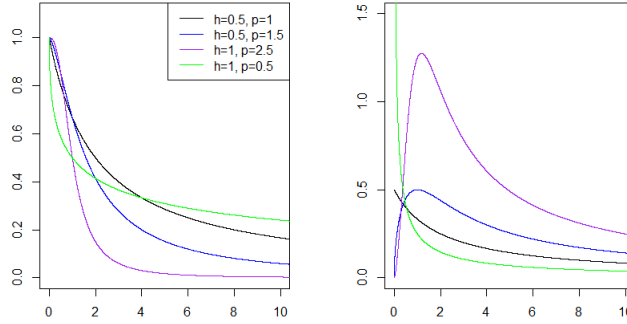


Figure 2.4: Survival and hazard functions along time, fixed the parameters

The parameters have here a central role and one of the most used way to fix them properly is using the maximum likelihood estimation. Anyway, if that is particularly hard, non-parametric models can be a better choice.

2.4 Non-parametric models

Useful for making simple comparison, for example during the exploration of data to have a general idea of its trend, non-parametric models are based on estimations that don't use distributional assumptions. We will focus only on the **Kaplan-Meyer** estimator for the survival function and the **Aalen-Nelson** estimator for the cumulative hazard.

Definition 2.4.1. Let N be the sample size, $t_1 < t_2 < \dots < t_{J-1} < t_J$ the ordered distinct failures (or event) time with $J \leq N$, d_j the number of failures registered at time t_j with $j \in \{1; \dots; J\}$, c_j the number of censored in the interval $(t_{j-1}, t_j]$ and n_1, n_2, \dots, n_j the number of patients at risk for each time registered, which is computed by the formula $n_j = N - \sum_{i=1}^j (d_i + c_i)$.

Then, the hazard can be estimated by:

$$\hat{\lambda}(t_j) = \frac{d_j}{n_j}$$

The Kaplan Meyer estimator can finally be defined as

$$\hat{S}(t) = \prod_{j|t(j) \leq t} \hat{p} = \prod_{j|t(j) \leq t} \frac{n_j - d_j}{n_j}$$

Note that, if we consider a time u when no new events or censoring are registered, $\hat{p}(u) = 1$.

From its definition, $\hat{S}(t)$ is a step function continuous to the right - i.e. $\hat{S}(t) = \hat{S}(t+)$ - ,

which decreases only at times of event.

The censoring affects only the "height" of the steps. Moreover, if nobody has been censored for an interval, let's say from t_1 to t_k , the formula can be simplified like the following: $\hat{S}_{t_k} = \frac{d_k}{n_1}$.

Pay attention to the number of people at risk: smaller it is, smaller the precision of the survival estimation.

Similarly as we derived the Kaplan-Meier estimator, we can obtain the cumulative hazard estimator:

Definition 2.4.2. With the same notation of 2.4.1, the Aalen-Nelson estimator is defined as

$$\Lambda(t) = \sum_{j|t_j \leq t} \frac{d_j}{n_j} = \sum_{j|t_j \leq t} \hat{\lambda}_j$$

Even Aalen-Nelson is a step function continuous to the right, but it is monotonic increasing and not decreasing like Kaplan-Meier.

2.5 Survival analysis main issues

To limit biases and avoid misleading results during a survival analysis, there is a list of warning points to pay attention to:

- **the definition of a proper starting point.** It should be precise and the same for every participants. For example, the arisen of a disease is vague and conditional to numerous factors that change for every subject, while the day of a surgical operation or of an hospitalization has no ambiguities. Whatever the choice will be, it will inevitably influences the rest of the study.
- **the definition of the event of interest.** It is not banal as it can superficially appear: first of all, it should be well-defined. For instance, does "death" refer to one specific cause of death or to an overall mortal event? Which happenings does the researcher identify as "relapse"?
Moreover, there is always the possibility that a *competing risk* obstructs the analysis: it is in fact an event that precludes the occurrence of the event of interest [9]. Sometimes, in that case, a possible solution is considering the *combined end-points*, which correspond to the first time of multiple considered events.
- **the modalities of follow-up**, whose updates should be done at regular intervals. The risk is indeed to focus with more attention to the bad happenings, not uniformly updating the data.
- **censoring and lost-to-follow-up.** Clearly, if the majority of patients is censored or lost during the study, this implies a significant problem. It's always necessary to control these percentages to check that the dataset doesn't contain any exaggerated

anomalies.

Second, the researchers should make considerations about removing or keeping data of subjects with a very limited period of follow-up, maybe deciding a minimum necessary period to be included in the analysis.

Finally, in survival analyses, the censoring must be **non-informative**, i.e. that censoring time must be independent from survival time ([8],pp.403-411). This assumption represents an important challenge for the clinical studies, since it is difficult to demonstrate.

2.6 Cox model

In the field of semi-parametric methods, Cox model is probably the most famous. It has been proposed by David Cox in 1972 in the paper "Regression models and life-tables" [10] and it consists in a regression model which relates time of event t and a set of explanatory variables \mathbf{X} .

In particular, this model is defined as:

$$\lambda(t) = \lambda_0(t) \exp(\beta X)$$

So, λ is given by a product between a function dependent only on time - λ_0 , called the **baseline hazard** - and a function dependent only on covariates, which are then considered time independent.

Defining λ in this way implies that we are also assuming **proportional hazards**:

$$\frac{\lambda_{X=1}(t)}{\lambda_{X=0}(t)} = \frac{\lambda_0(t) \exp(\beta)}{\lambda_0(t)} = \exp(\beta)$$

And more in general:

$$\frac{\lambda_{X=x+1}(t)}{\lambda_{X=x}(t)} = \frac{\lambda_0(t) \exp(\beta(x+1))}{\lambda_0(t) \exp(\beta x)} = \exp(\beta)$$

In other words, the **hazard ratio (HR)**, which is simply a quotient between two hazards, is constant in time.

Let $X = 0$ and $X = 1$ indicate respectively the unexposed and the exposed subject. Given the hazard ratio between $\lambda_{X=1}$ and $\lambda_{X=0}$, $\mathbf{HR}_{1,0} > 1$ indicates that the exposure considered is a risk factor, while it is a protection if $\mathbf{HR}_{1,0} < 1$. It seems, instead, not to have influence if $\mathbf{HR}_{1,0} \sim 1$.

In light of this, the hazard ratio represents a quantity of a great interest to estimate during the analysis.

Let's now see how to construct the model.

2.6.1 Parameters estimation

We defined the structure of a Cox model and we saw that, to be complete, it needs the parameters β to be defined.

For this purpose, the parameters are estimated using the maximum likelihood method utilizing the *partial* likelihood function, i.e. a formula that considers probabilities only of subjects who had the event and not of those ones who had been censored ([7], ch. 3.2).

Definition 2.6.1. Let N be the sample size, $t_1 < t_2 < \dots < t_{J-1} < t_J$ the ordered distinct failures time with $J \leq N$, \mathbf{x}_i with $i \in \{1, \dots, N\}$ a covariate, $R(t_j) = \{i | t_i \geq t_j\}$ the risk set - i.e. the set of subjects at risk at time t_j -. Then, the partial likelihood function is defined as

$$pl(\beta | \text{data}) = \prod_{j=1}^J \frac{\lambda_{\mathbf{x}_j}(t_j)}{\sum_{i \in R(t_j)} \lambda_{\mathbf{x}_i}(t_j)} = \prod_{j=1}^J \frac{\exp(\beta' \mathbf{x}_j)}{\sum_{i \in R(t_j)} \exp(\beta' \mathbf{x}_i)}$$

Each factor of the formula corresponds to the probability that a subject with covariate \mathbf{x}_j has the event at time t_j , given the risk set $R(t_j)$ [11]. It's important to notice that pl depends only on β and not on the baseline hazard λ_0 or on time.

The point of maximum is the same for pl and $\ln(pl)$, since the logarithm is monotonous increasing, so we calculate the log-partial likelihood:

$$\ln(pl(\beta | \text{data})) = \sum_{j=1}^J \beta' \mathbf{x}_j - \sum_{i \in R(t_j)} \ln(\exp(\beta' \mathbf{x}_i))$$

Finally, we want to find the zeros of the score function $\mathbf{U} = (U_1, \dots, U_N)$ - which is the gradient of the log-likelihood function with respect to the parameter vector -, where U components are equal to:

$$U_k(\beta) = \frac{\partial \ln(pl(\beta | \text{data}))}{\partial \beta_k} = \sum_{j=1}^J \left(x_{k,j} - \frac{\sum_{i \in R(t_j)} x_{k,i} \exp(\beta' \mathbf{x}_i)}{\sum_{i \in R(t_j)} \exp(\beta' \mathbf{x}_i)} \right)$$

$x_{k,j}$ indicates the k covariate for the subject who has an event at time t_j , while the ratio is the weighted average of the covariate x_k on risk set at time t_j .

2.7 Log-rank test

The survival analysis purpose is to compare survival distributions of two (or more) treatment groups. For this reason, tests of significance are utilised to verify if there is a considerable difference in the effects of the studied exposures.

One of the most popular is the **Log-rank test** LR , whose null hypothesis is that $\lambda_A(t) = \lambda_B(t)$, where A and B indicate the treatments [12].

Definition 2.7.1. Consider the set of failures registered at time t_j for the group of people treated with A which have been *observed* - O_{Aj} -, then the set of failures that have been *expected* - E_{Aj} . The latter is equal to $E_{Aj} = \frac{d_j}{n_j} n_{Aj}$, i.e. it is the ratio between the total number of failures and the number of subjects at risk at time t_j , multiplied by

the number of subjects treated with A and at risk at t_j .
The log-rank is based on $O_{Aj} - E_{Aj}$, in particular it is given by

$$LR = \frac{(O_A - E_A)^2}{Var(O_A)}$$

The higher LR , the smaller is the probability that the data is consistent with H_0 .
Note that, under H_0 , $LR \sim N(0, 1)$ and $LR^2 \sim \chi^2$: these approximations are used to test the null hypothesis.

The log-rank test major efficiency is reached with proportional hazards. To fit more generic cases, some extensions of the test have been thought, such as the weighted and the stratified log-rank tests.

Finally, note that LR indicates only if there is a difference in the survival distribution of treatment groups, but doesn't say anything about the size of this difference. To have estimations of the dimension of treatment effects, hazard ratio is commonly considered [13].

Other tests of significance

While the log-rank test is specifically built for the survival analysis, other tests of significance commonly utilised are the **Wald test**, the **score test** and the **likelihood-ratio test**, which are generically used in inferential statistical studies [14].

Their null hypothesis is $H_0 : \beta = 0$, for a generic parameter β . In our case, it refers to one of the Cox model parameters.

An important property of the score test is the fact that it is equivalent to the log-rank test with univariate Cox models. The likelihood-ratio test, instead, is asymptotically equivalent to LR under the proportional hazards assumption.

2.8 Estimations of treatment effects

Let $Y_i(0)$ indicate the outcome of the i -th patient when they didn't receive the treatment and $Y_i(1)$ the outcome when they received it. Then, the treatment effect for the subject i can be expressed as $Y_i(1) - Y_i(0)$. How can we estimate it? In fact, it seems that it's impossible to compute both $Y_i(1)$ and $Y_i(0)$, an issue that is even known as the fundamental problem of causal inference [15].

Definition 2.8.1. It is impossible to *observe* the treatment effect, because it's not possible to *observe* $Y_i(1)$ and $Y_i(0)$ at the same time, since the subject i or is treated or is not, not both. This issue is called **the fundamental problem of causal inference**.

Although one could think that this problem would invalid the whole theory of causal inference, there are two possible solutions, that Holland named in [15] as **the scientific solution** and **the statistical solution**.

The first resolution is very simple to imagine: we will compute two experiments A and B , A that gives $Y_i(1)$ and B that returns $Y_i(0)$. To be valid, there must be no differences

between these two trials, except the value of Z and the outcome. In other words, we are making an invariance assumption, such that, if we had measured $Y_{i,A}(1)$ in A and then again $Y_{i,B}(1)$ in B instead of observing outcomes for different values of Z , the outcomes would have coincided ($Y_{i,A}(1) = Y_{i,B}(1)$).

It is pretty evident that the researcher won't ever be able to completely demonstrate this invariability, so, an act of faith is partially necessary. Moreover, there are also practical concerns in pursuing this approach.

What really interests us is the statistical solution, which is even described in Rubin's potential outcome framework [16]. Consider $Y_i(1)$ and $Y_i(0)$ as the two *potential* outcomes each subject i can have, not strictly the observed ones. Then, we can define the average treatment effect ATE and the average treatment effect on the treated ATT [17]:

Definition 2.8.2. The **average treatment effect** ATE refers to the average effect of moving an entire population from untreated to treated.

In particular,

$$\text{ATE} = E[Y(1) - Y(0)]$$

Definition 2.8.3. The **average treatment effect on the treated** ATT refers to the average effect of treatment only on subjects who actually received the treatment.

Let Z be the variable which indicates if the patient is treated ($Z = 1$) or not ($Z = 0$). Then,

$$\text{ATT} = E[Y(1) - Y(0)|Z = 1]$$

At the end, each patient either receive or not the treatment, but in this way it's possible to quantify the treatment effect.

Note that in random clinical trials, these two quantities should coincide, because control and cases groups are chosen to be similar. In observational studies, on the contrary, the investigator have to decide which quantity estimate, because they are not equivalent. This choice depends on the research question, so, for example, if we are comparing two therapies for the same disease, it's likely that we would estimate ATE, because all patients could easily receive either treatment. On the other hand, the ATT is probably more appropriate for cases where not all the eligible patients are likely to elect for receiving the treatment [18].

In RCTs, the estimation of ATE can be done directly from the data, thanks to the randomization [19]: $\text{ATE} = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]$. With continuous outcomes, this formula corresponds to a difference of mean, while, with binary outcomes, it is calculated as a difference in proportion or an absolute risk reduction.

In observational studies, the absence of randomization unfortunately implies the presence of confounding factors that influence the outcomes, so the estimation of treatment effects becomes harder. Nevertheless, a solution has been found and it consists in the **propensity score methods**.

However, before introducing them, it is necessary to distinct two other quantities: the conditional treatment effect and the marginal treatment effect [18].

Definition 2.8.4. The **conditional treatment effect** is the average effect that is obtained switching one patient's status from untreated to treated, at an individual level.

Definition 2.8.5. The **marginal treatment effect** is the mean difference in outcomes of two identical population, except for the fact that one is composed only by treated subjects and the other by untreated.

There is the possibility that these quantities coincide: in that cases, the measure of the treatment are said to be **collapsible**.

Note that, in general, for a population there is only a single value of marginal treatment effect, but many of conditional treatment effect, in particular one for each set of covariates included in the regression model.

During survival analysis, it's fundamental to stabilize which effect is of our interest and, consequently, decide a suitable method and measure [20].

2.9 Propensity score methods

The problem of estimating causal effects in observational studies is clearly explained by the **Simpson's paradox** [21].

Definition 2.9.1. The Simpson's paradox is defined as a statistical phenomenon where a trend or a result is observed in a population, but it reverses or disappears when the population is divided in subpopulations.

To explain this paradox, the UC Berkley's suspected gender-bias example is usually utilised. In 1973, the school rejected the 65% circa of female applications, versus the 56% of male ones: someone can then erroneously conclude that women had less probabilities to be admitted in that university. Actually, analysing deeper the data, for the majority of the departments, there was a statistically significant gender bias *in favour of females* - and no gender bias in the rest of the departments. How did they explain this result apparently absurd? It was found out that women tended to apply to the most selective departments: this hidden variable brought to misleading results when the whole data was considered, reversing the actual trend.

We risk to fall into the same paradox during the analysis of observational study data. In fact, the treatment groups that are considered are not homogeneous, due to the non-randomized nature of these studies, and so, there are various confounding factors and hidden influencing variables. To face these obstacles to the analysis, the propensity score methods have been developed [19].

Definition 2.9.2. Given the observed covariates X , the **propensity score** e is the probability of treatment assignment conditional on X :

$$e(X) = P(Z = 1|X)$$

It is a *balancing score*, i.e. a function such that the conditional distribution of X given e is the same for exposed and unexposed subjects [22].

This function can be applied to observational studies to mimic some characteristics of

a randomized trial, to permit the estimations of causal effects. In particular, the essential property of RCTs that is desired to obtain is the **strongly ignorable treatment assignment**, as it is named by Rosenbaum and Rubin in [22]:

$$(Y(1), Y(0)) \perp Z|X \quad \& \quad 0 < P(Z = 1|X) < 1$$

The outcome $(Y(1), Y(0))$ is conditionally independent from the assignment of the treatment Z , given the covariates X , and every subject of the studied population has a chance of receiving the treatment.

For our case of study, we are going to use two of the most popular propensity score methods: **Matching** and **Inverse probability weighting** (IPW).

Anyway, we need to study the causal effects of three treatments, not only two. Note that, so far, we only focused on definitions and explanations considering binary exposure: with multiple treatments, the basic knowledge is the same, but the analysis gets harder, requesting additional assumptions and modelling techniques [23]. For this reason, we are going to utilise also the **generalised propensity score cumulative distribution function method** (GPS-CDF), proposed in [24].

Since regression models are requested during the construction of these methods, we will discuss at the end of the chapter about the two models we've chosen to utilise - multinomial logistic regression and generalized boosted model.

2.9.1 Matching

The **propensity score matching** method consists in forming matched sets of treated and untreated subjects, that have similar propensity score values - so, consequently, similar baseline covariates distributions -.

There are various possible ways to construct the matched sets. First of all, it is necessary to stabilize the ratio between treated and untreated patients in the sets: we are going to use the *1:1* or *pair matching*, which is also the most popular, but other two possibilities are *many to one matching* (many untreated vs one treated) and *full matching*, where sets can contain one treated and at least one untreated or vice versa.

Before forming the matched sets, we need to decide about:

- **Matching without replacement** or **matching with replacement**. In the first scenario, each untreated subject is associated with at maximum one treated individual. If we are comparing two different treatments, there is then the possibility that some patients can be discarded from the analysis, if people that received the other treatment have already been all matched. This happens when the total number of subjects treated with A is lower than the total number of subjects treated with B, or vice versa.
By contrast, in matching with replacement, the untreated subjects can be matched with more than one individual.
- **Greedy matching** or **optimal matching**. With the first approach, each treated subject, selected randomly time to time, is matched with the untreated subject

that has the nearest propensity score.

Optimal matching, in contrast, aims to minimize the total within-pair difference of propensity score values. Anyway, it has been shown that the latter isn't better in balancing baseline covariates than the greedy approach [25].

- **Distance criteria.** There are many possible ways to define distances (or similarities). However, in this context, two main methods are considered. The first is nearest neighbour matching: the treated subject is associated with the untreated individual who has the closest propensity score. If more untreated patients are equidistant to the same treated person, one of them is randomly selected. The second approach is analogous, but a caliper distance is fixed: absolute difference of propensity scores that belong to the treated and untreated matched patients must be lower than the caliper. If it is higher, the match is discarded and so, even the unmatched treated subjects.

To decide which approach utilise, it is necessary to make a trade off between precision, computational costs and matched sample size and choose what is more important for the study ¹.

Note that the matching model can be utilised only to estimate the ATT and not ATE. To computed the matched sets, we are going to utilise the R package *MatchIt* [27].

2.9.2 Inverse probability weighting

The main issue of a non-randomized study is the unbalanced distribution of baseline covariates in the treatments groups. Thus, it was born the idea of introducing weights to equilibrate the features: this is the **inverse probability weighting using the propensity score** method.

The weight of the i -th subject is defined as

$$w_i = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}$$

i.e. it is the reciprocal of a patient's probability of receiving the treatment that they *actually* received. At this point, the regression models, for example the Cox model, can be weighted to estimate the treatment effect.

Anyway, often a problem arises: some extreme weights could be generated, when some patients have very low probabilities of being subjected to the treatment they actually been subjected to. In that cases, a few subjects have a great importance in the model. To limit this inconvenience, two strategies are used: the adoption of **stabilized** weights and the introduction of a threshold to trim outlier weights. In particular, the i -th subject's stabilized weight can be defined as

$$w_i = \frac{Z_i p_i}{e_i} + \frac{(1 - Z_i)(1 - p_i)}{1 - e_i}$$

¹For an extensive comparison of matching methods, see Austin's paper of 2014 [26].

where p_i is the probability of receiving the treatment without considering the covariates, i.e. it is equal to $\frac{\# \text{ patients subject to the treatment}}{\# \text{ overall patients}}$ [28].

One advantage of the IPW model is the fact that it can be used to estimate both ATE and ATT. In particular, the weights described before allow to estimate the average treatment effect, but if we consider

$$w_i = Z_i + \frac{(1 - Z_i)e_i}{1 - e_i}$$

we can estimate the ATT.

We are going to compute the weights utilising the R package *WeightIt* [29].

2.9.3 Generalised propensity score cumulative distribution function

Causal inference applied to the survival analysis of multiple treatments requests extensions and generalisations of the definitions and methods utilized with binary exposures. However, in scientific literature, there are few papers about this topic.

As already said, we've decided to apply to our case of study the matching propensity score method and the IPW method, adapted in a simple way to the multi-treatment setting. Moreover, we are going to utilise also a method created ad-hoc for the latter setting: the **generalised propensity score cumulative distribution function** model, proposed in [24].

Let's first start with some definitions.

Multi-treatment setting: definitions

Usually, treatments are considered to be binary, but they can be even categorical, ordinal or continuous. For this reason, it has been developed the concept of generalised propensity score [30]. Anyway, in this work, we will refer with GPS only to the propensity score adjusted for the multi-treatment setting.

Definition 2.9.3. The **generalised propensity score** (GPS) of the i -th subject is defined as a vector of n_t components, where n_t corresponds to the number of the considered treatments, such as

$$e_i(j|\mathbf{X}) = P(Z_i = j|\mathbf{X}) \quad \wedge \quad j \in \{0, \dots, n_t - 1\}$$

The letter j indicates the treatment considered.

The IPW method that we are going to use simply extends to the multi-treatment setting thanks to this new definition of propensity score. The problem of this choice is the fact that it considers only one entry of the GPS vector. Note that higher is n_t , less information is contained in a single component of GPS. However, since in our case of study we compare only three treatments, the method should still result as convenient, as it has been shown in the simulation study contained in [31].

About the average treatment effect, its definition is simply extended such that

$$\text{ATE}_{j,j'} = E[Y(j) - Y(j')]$$

Consequently, we have $2n_t$ values of ATE. Having so many values to consider can be problematic or, at least, annoying. For this reason, if it's possible to define an order in the treatments set, it's preferable to fix a "treatment zero" t_0 and to estimate only the ATE_{j,t_0} for the rest of the treatments j . In this way, the number of ATEs that we are considering is reduced to $n_t - 1$, even if, the existing ATEs are still $2n_t$. Note that similar considerations are valid for the average treatment effect on the treated, which is now defined as:

$$ATT_{j,j'} = E[Y(j') - Y(j)|Z = j]$$

ATT has the advantage that there is a reference treatment compared to all the others from its proper definition, so there are only $n_t - 1$ values of ATT.

In our case of study, the semi-anatomic resection technique is the middle way between the other two surgeries. Therefore, we decided to consider the anatomic resection as the "reference treatment" and compare the other two techniques to it. Anyway, I will discuss this aspect in the application chapter 3.

Development of the model

We introduced the concepts that regard generically the multi-treatment survival analyses. Now, we can focus on the GPS-CDF method.

Thanks to how we defined the generalised propensity score, we can consider it as a discrete probability distribution and, then, create a *probability mass function* PM [32].

Definition 2.9.4. A **probability mass function** PM indicates the probability that a discrete random variable assumes a particular value.

Let Z_i be a variable that takes values in \mathbb{Z} and $j \in \mathbb{Z}$,

$$PM(j) = P(Z_i = j)$$

In particular, the probability mass function derived from GPS corresponds to $PM_i(j) = e_i(j|X)$. Note that each subject has an associated PM function, whose shape can be compared to other PM shapes: similar shapes indicate close GPS vectors, and, accordingly, close baseline covariates distributions.

Anyway, a function easier to analyse and more suitable for our purpose is the cumulative distribution function CDF ([33], p. 178).

Definition 2.9.5. A **cumulative distribution function** is a function $CDF : \mathbb{R} \rightarrow [0, 1]$ that indicates the probability of a variable Z of taking a value smaller or equal to z :

$$CDF(z) = P(Z \leq z)$$

Our objective is to derive one significant value for each subject, to be able to compare patients' similitude. The idea is to define a one parameter function starting from CDF , such that patients with nearer parameters should have nearer GPS. Note that, like with PM , subjects with more similar CDF shapes have closer GPS vectors.

We know that it is possible to build a one parameter function that will fit *CDF* shape, since it is monotone increasing and assumes values in $[0, 1]$. In particular, we define the parameter function g as a power function such that

$$g_{i,\alpha}(z) = h_{i,z}^{exp(\alpha)} \approx CDF_i(z)$$

where α is the parameter and z indicates the treatment. In the formula, $h_{i,z}$ is a standardized value associated to each treatment.

The final step is fitting g to *CDF*: we utilise a non-linear least squares algorithm (NLS). The mean distance between the two functions is minimized in respect to α :

$$\min_{\alpha} \sum_{z=0}^{n_t-1} (h_{i,z}^{exp(\alpha)} - CDF_i(z))^2$$

We have finally obtained the scalar values that will permit us to estimate the similitude between subjects.

In particular, we utilize the matching method with parameters α instead of propensity scores. In other words, we match subjects with the lowest mean absolute difference of the correspondent parameters. The method is the same described before.

Note that the order of treatments in their GPS vector influences the correspondent *CDF* function. Therefore, if they aren't ordinal, it is necessary to compute all the possible *CDF* functions and the associated parameters and, at last, to select the order that gives the best covariates balance.

Anyway, in our case of study, treatments can be ordered, so we will develop the *GPS – CDF* model following only the order "anatomic"- "semi-anatomic"- "wedge".

To implement this model, we are going to utilise the *GPSCDFR* R package [34], created by part of the authors of [24].

2.10 Models for the estimation of propensity score

The core description of the models that we are going to utilize to estimate the causal treatment effects has already been illustrated. However, we didn't focus on how to estimate the propensity score values, a fundamental step to complete the construction of the propensity score models.

This estimation is realized thanks to classification models which have as outcome variable the treatment and as explanatory variables the baseline covariates. In particular, we decided to rely on the multinomial logistic regression (which is also the default model utilised by many R functions) and on the generalised boosted regression (that corresponds to the R function *gbm* [35]).

2.10.1 Multinomial logistic regression

The **multinomial logistic regression** is the extension of the binary logistic regression for multiple outcomes. Despite its name, it is a classification model.

Let's start with the binary case. The name "regression" is not casual: the idea of the method is to build a linear regression of the log-odds on the baseline covariates.

Definition 2.10.1. The **log-odds** is the logarithm of the odds of an event, where

$$odds(E) = \frac{P(\text{success of } E)}{P(\text{failure of } E)} = \frac{p}{1-p}$$

and p is the probability of success of the event E .

Clarified that the odds is connected to the probability of an event [36], this regression permits, with some simple steps, to estimate the probability of receiving a certain treatment. The logistic classifier, then, returns for each patient the treatment t with the highest probability to be assigned [37].

Now, let's examine the regression steps [38]: as already said, we search for an equation between log-odds and baseline covariates such that

$$\log \left(\frac{p(X)}{1-p(X)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_N x_N = \boldsymbol{\beta}' X \quad (2.2)$$

where $p(X)$ refers to the probability of receiving the treatment for a subject i with baseline covariates X , N is the number of baseline covariates and $\boldsymbol{\beta}$ is the vector of regression coefficients (the index i is omitted from the formula to make the notation lighter).

From 2.2, we obtain the following equation, applying the exponentiation to the both sides:

$$\frac{1}{1-p(X)} = \exp(\boldsymbol{\beta}' X)$$

Then, we can easily obtain:

$$p(X) = 1 - \frac{1}{\exp(\boldsymbol{\beta}' X) + 1} = \frac{\exp(\boldsymbol{\beta}' X)}{1 + \exp(\boldsymbol{\beta}' X)} = \frac{1}{\frac{1+\exp(\boldsymbol{\beta}' X)}{\exp(\boldsymbol{\beta}' X)}} = \frac{1}{1 + \exp(-\boldsymbol{\beta}' X)} = S(X) \quad (2.3)$$

where S is the sigmoid function.

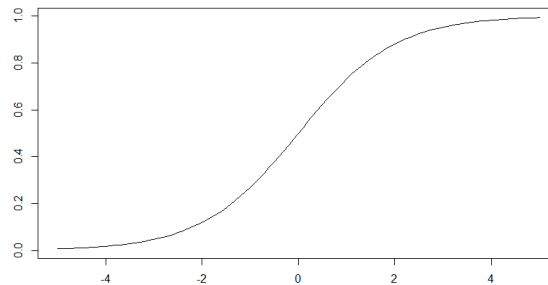


Figure 2.5: Sigmoid function

Now, we can simply estimate β with the maximum log-likelihood method. The extension to the multinomial case is simple [39]: we define a "treatment zero" t_0 , such that the odds of each treatment t will be considered equal to $\frac{p_t}{p_{t_0}}$, where p_t is the probability that the patient i receives the treatment t , while p_{t_0} the probability that they receive the treatment t_0 .

In this way, the equation 2.2 becomes

$$\sum_{l=1}^{n_t-1} \log\left(\frac{p_t(X)}{1 - p_{t_0}(X)}\right) = \beta'_l X$$

where n_t is the total number of the considered treatments and there are $n_t - 1$ regression coefficients vectors β , i.e. one vector for each treatment except t_0 .

So, in the end, the equation 2.3 is generalised as

$$p(X) = 1 - \frac{1}{1 + \sum_{l=1}^{n_t-1} \exp(-\beta'_l X)}$$

The estimations of the propensity scores and of the generalised propensity score vectors are finally obtained by the estimated values of $p(X)$.

We implemented this model with the R function "multinom" from the *nnet* package [40] for the GPS-CDF method, while, for the rest of the propensity score functions, the multinomial logistic regression is the default option.

2.10.2 Generalised boosted regression

The **generalised boosted regression** is a model obtained by the fusion of regression trees and the boosting method [41]. It follows the idea proposed in the Gradient Boosting Machine [42], as it is reported in the *gbm* package [35].

First of all, a regression tree - which is simply a decision tree that can be used even with continuous variables as target variables - predicts the probability of receiving a certain treatment for each patient i (given their covariates X_i). This probability coincides with the propensity score we want to estimate.

More specifically, this tree uses an iterative process, where the dataset is splitting into subsets until a chosen loss function is minimized.

Then, the obtained result is given in input to a boosting method, which is an ensemble learning technique that combines different models that singularly work poorly, while together formed a stronger learner [43]. In particular, the data that was badly modelled by one tree has more probabilities to be selected by the following tree, with the purpose of enhancing the model in every step.

The advantage of the generalised boosted regression is its effectiveness with non-linear treatment models, but at the same time it performs worse with simpler models [44].

3. Case of study

3.1 Clinical context

In 2020, the liver tumour caused 830000 deaths, positioning as the **third** most mortal cancer worldwide [45]. Moreover, it has been estimated that, by 2025, there will be more than one million of new cases *per year* [46]. In particular, the 90% of the cases of this global health challenge consists in the hepatocellular carcinoma (HCC).

The liver cancer risk factors are well known and the most common ones are the chronic viral hepatitis (both HBV and HCV), which unfortunately are the main causes of the tumour diffusion due to their infectivity [47]. In particular, they lead to cirrhosis, which is a liver disease that is present in the majority of patients affected by liver neoplasia. Cirrhosis can be caused also by other factors, such as the abuse of alcohol, the non-alcoholic fatty liver disease - which is widespread in obese people - and some types of liver autoimmune diseases such as the primary biliary cirrhosis.

Another risk factor responsible of many cases of liver tumours in low-income countries is the aflatoxin, a mycotoxin generated by the fungi *Aspergillus flavus* and *Aspergillus parasiticus*, that can be ingested through contaminate food [48].

Prevention programs are realized to limit the diffusion of these risk factors. In addition to them, it is necessary to continue investing in the research to find more effective cure. Unfortunately, so far, the therapies usually adopted have to face many obstacles.

The most efficient treatment for the hepatocellular carcinoma is the liver transplantation, which is nevertheless very difficult to realize, because of the very selective criteria for the transplantation eligibility, the scarce availability of organs and the high costs of the intervention. Therefore, the first-line therapeutic option is the **surgical resection**, which can still be operated only in the 30% circa of the patients. The surgery has also warning postoperative long term recurrence rate: the 50% during the first three years and >70% during the first five years [49]. For this reason, this study wants to examine the relationship between the resection techniques and the time of death or relapse. Is the choice of the surgical intervention critical for the relapse-free and the overall survival?

To answer this question and investigate better the problem of HCC recurrence and mortality, in 2018, the project named HE.RC.O.LE.S. [50] was set up: its objective is making available to the oncological research an HCC Italian Registry, i.e. a network of Italian centres which share the data about their patients affected to HCC and who were operated.

This data keeps track of patients' characteristics and of recurrence risk factors, which have been identified as HCC dimension, grading, microvascular invasion and satellitosis. Moreover, a feature is dedicated to the resection techniques, that we have decided to distinguish into three categories:

1. the **anatomic** resection, which follows the *Brisbane division*. It is a classification

of the liver that divides it in nine segments, each one with unique blood supply and biliary drainage [51] (see the figure 3.1). It is preferred with malignant tumours, since it cuts with large margins from the cancerous mass to have the certainty of removing it all.

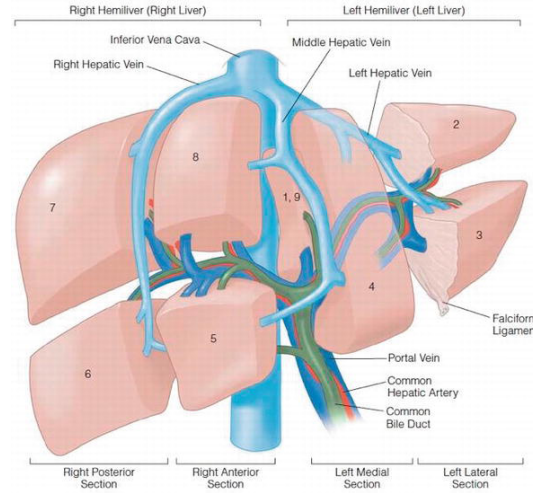


Figure 3.1: The Brisbane division of the liver anatomy, from [51]

2. the **wedge** or non-anatomic resection, which is more conservative and doesn't follow the segmental planes, with the intent of preserve the organ as much as possible. It is mostly used with benign tumours, or, by contrast, with critical and vast forms of cancer, when it's not possible removing the whole malignant mass without putting the patient's life in danger.
3. the **semi-anatomic** resection, which is the middle way.

Note that, during the data collection, there are criteria for inclusion and exclusion of the patients. In particular, the subjects included must have a confirmed diagnosis of hepatocellular carcinoma and must have been treated for that with a surgical operation in a study participating center, in a period included between 2008 and 2017. By contrast, the subjects have been excluded if the surgery was a downstaging procedure for transplant or if they have combined liver primary neoplasms, had been affected by other tumours in the past, or had been previously undergone to a liver transplantation. There is no age limit.

Obviously, all the data has been anonymized.

Actually, we are examining only a subset of this data, such that the patients considered are of age, presented large hepatocarcinoma (its size must be bigger than 3.5 cm), had solitary nodule cancers and hadn't been subjected to any therapies before the surgery. The object of our study is estimating the marginal treatments effects, verifying if one of the resection techniques results to be better than the others to cure large HCC.

3.2 Dataset

We now focus on the exploration of the data considered in this thesis.

The dataset is composed by 613 instances - i.e. 613 patients - and 75 features. We won't describe all of them, since we make an initial selection, discarding all the variables that are evaluated as not interesting for this case of study. To sum up, the excluded features are:

- variables that have been used to compute other variables or, more generally, that are strictly correlated to other features;
- variables that are almost homogeneous. Some of them have been used to select the patients who present large cancers: remember that we consider only solitary nodule cancers, with size bigger than 3.5 cm.
- variables that are linked to post-surgery conditions, so that are not diagnostic factors.

What remains is 28 features, whose statistical characteristics are summarized in the tables 3.2 and 3.3.

















	Mean	SD	Min	Median	Max	NA	Boxplot	Histogram
timeOSmesi	36.76	32.77	0.00	27.38	145.61	30		
timeRFSmesi	27.93	30.15	0.00	15.87	145.61	25		
Age	67.66	12.77	0.10	70.83	88.62	1		
Bilirubin	0.91	0.63	0.04	0.78	8.96	21		
Albuminemia	3.95	0.53	1.70	4.00	5.40	64		
Creatinine	0.95	0.33	0.22	0.90	3.00	70		
Platelets	215.25	98.32	44.00	197.50	721.00	23		
INR	25.57	13.43	1.00	23.00	73.00	31		

Figure 3.2: Overall descriptive statistics of numerical variables

eventomorte (%)	RFS (%)	PROCEDURE (%)	Sex (%)	ASA (%)	CCI (%)	MELD_score (%)	Cirrhosis (%)	child_pugh_grade (%)
0 403 (65.7)	0 262 (42.7)	0 167 (27.2)	Male	473 (77.2)	1 16 (2.6)	4 2 (0.3)	0 349 (56.9)	0 332 (54.2)
1 198 (32.3)	1 139 (55.3)	1 160 (26.1)	Female	140 (22.8)	2 177 (28.9)	2 12 (2.0)	1 257 (41.9)	1 246 (40.1)
NA	NA	2 43 (7.0)		3 270 (44.0)	3 36 (5.9)	7 137 (22.3)	NA	2 18 (2.9)
		3 28 (4.6)		4 13 (2.1)	4 45 (7.3)	8 94 (15.3)		NA
		4 170 (27.7)		5 2 (0.3)	5 82 (13.4)	9 63 (10.3)		
		5 44 (7.2)		NA	6 88 (14.4)	10 39 (6.4)		
		NA			7 99 (16.2)	11 25 (4.1)		
					8 59 (9.6)	12 9 (1.5)		
					9 24 (3.9)	13 12 (2.0)		
					10 16 (2.6)	14 4 (0.7)		
					11 2 (0.3)	15 2 (0.3)		
					12 3 (0.5)	16 4 (0.7)		
					NA	146 (23.8)		
						19 2 (0.3)		
						21 1 (0.2)		
						25 1 (0.2)		
						27 1 (0.2)		
						NA		
						114 (18.6)		

eventomorte (%)	HCV (%)	HBV (%)	Potus (%)	inv_macro_port (%)	inv_macro_svrapp (%)	EdmondsonGrading (%)	MVI (%)	Satellitosis (%)	Capsule (%)	R (%)
0 428 (69.8)	0 408 (66.6)	0 491 (80.1)	0 480 (79.9)	0 509 (83.0)	0 477 (77.8)	1 50 (8.2)	1 50 (8.2)	5 (0.8)	0 406 (66.2)	0 275 (44.9)
1 163 (26.6)	1 192 (31.3)	1 109 (17.8)	1 104 (17.0)	1 78 (12.7)	1 56 (9.1)	2 277 (45.2)	2 277 (45.2)	0 247 (40.3)	1 111 (18.1)	1 196 (32.0)
NA	NA	NA	NA	26 (4.2)	80 (13.1)	3 227 (37.0)	3 227 (37.0)	1 304 (49.6)	96 (15.7)	1 142 (23.2)
						4 33 (5.4)	4 33 (5.4)	2 1 (0.2)	NA	2 7 (1.1)
						NA	26 (4.2)	NA	56 (9.1)	NA
										80 (13.1)

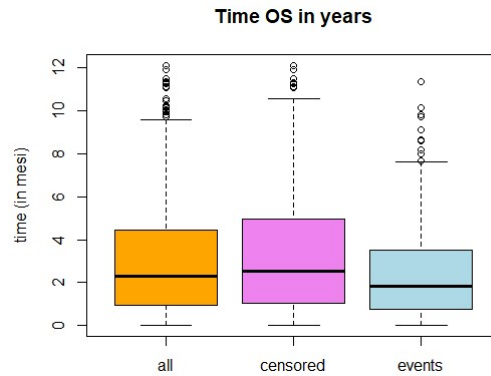
Figure 3.3: Overall descriptive statistics of categorical variables

Output Variables

1. *eventomorte*. A binary variable that indicates if the patient died (1 = death, 0 = alive).

The death is registered for the one third circa of the patients.

2. *timeOSmesi*. The number of months associated to eventomorte: it indicates the period of follow up.

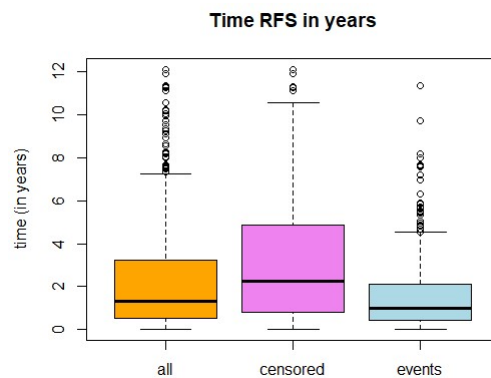


Note that the minimum value is 0: this extreme case can cause problem. For this reason, we substitute it with $1/30$, that corresponds to one day. In this way, it is still very small, but we avoid problems that could emerge with null values.

3. *RFS* (Relapse Free Survival). A binary variable that indicates if there is a relapse (1 = relapse, 0 = alive).

More than the half of the patients has a relapse. Obviously, this event includes also the death.

4. *timeRFSmesi*. The number of months associated to RFS: it indicates the period of follow up.



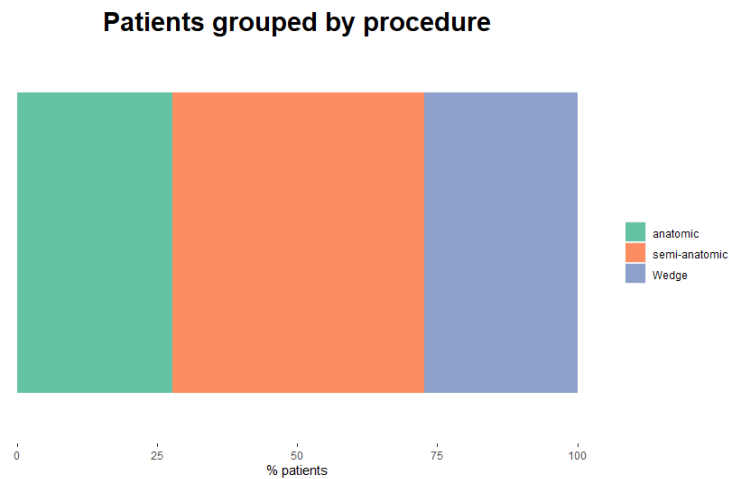
Like with *timeOSmesi*, the minimum value is 0. We again substitute this value with $1/30$.

Note that, for both OS and RFS, the times of censoring are averagely longer than the times of event: this is important, because, otherwise, there is a high risk of censoring a little before the happening of the event. In other words, it is necessary to follow up the patients for a reasonable duration.

Treatment Variable

5. *Procedure*. A number from 0 to 5, where 0 indicates the wedge resection, 4 the anatomic resection, 1 to 3 the semi-anatomic resection and, finally, 5 corresponds to the category "other techniques".

We have decided to discard the instances equal to 5. In fact, this resection technique label is too vague to be considered.



The surgery method most used is semi-anatomic resection. Anyway, the other two types of operation have been utilised with more than the one fourth of the patients.

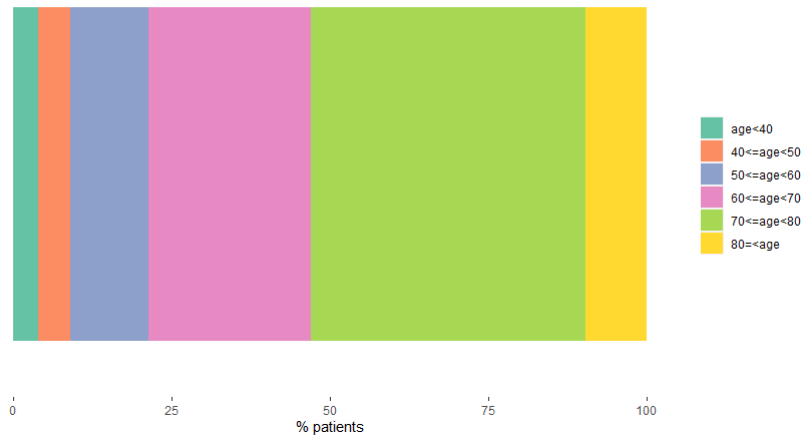
Anagraphic Data

6. *Age*.

We removed 3 instances with age lower than 18: minors shouldn't be included in the study.

In general, older the participants, higher is the risk of death or relapse.

Patients grouped by age

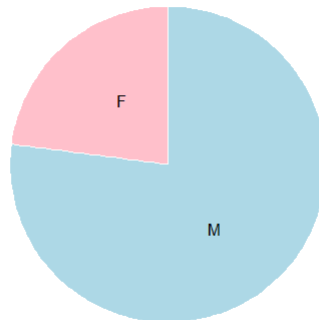


The most numerous group of patients has an age included between 70 and 80. Moreover, the great majority of the patients is older than 60.

7. *Sex*.

It has been documented that the hepatocarcinoma has significant higher incidence in males [52].

% patients divided in respect to sex



Even our dataset presents a considerable difference between the number of men and the number of women: circa the 2/3 of the patients is a male. This is the only considered feature without missing values.

Clinical Details

8. *ASA*. It is the acronym of American Society of Anesthesiologists [53] and assumes values from 1 to 5, numbers that indicate the patient's status before the anesthesia:

higher the value, worse the condition.

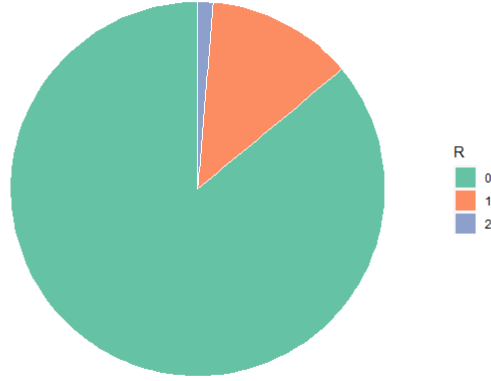
The majority of the subjects belong to the categories 2 or 3.

Unfortunately, there are numerous null values, precisely 135.

9. *CCI*. It is the acronym of Charlson Comorbidity Index [54]: high values mean that the patients already presents some pathologies and have a high mortality risk. It assumes values from 0 to 12.
Most of the subjects is classified with the values 5, 6 or 7, in the middle of the scale.
Like for *ASA*, this feature presents many missing values, too: 146.
10. *Meld Score*. It stages for model for end-stage liver disease [55]. It's an index about the liver functionalities that here assumes values from 4 to 27: higher is the number, worse is the situation of the organ.
Almost the totality of the subjects has a score lower than 13. Unfortunately, there are 114 missing values.
11. *Cirrhosis*. It is a binary variable that indicates if the patient's liver is affected to cirrhosis (1=with cirrhosis, 0=without cirrhosis).
Most of the participants don't have it, still the percentage of cirrhotic subjects is $\sim 42\%$.
12. *Child Pugh Grade*. It indicates the severity of the liver disease, classifying the patient's conditions with three levels (0,1,2): higher the level, worse the situation.
The subjects who belong to the second category represent a strict minority. Grade 0 participants are more than the half of the total, while grade 1 participants correspond to the 40% of the total circa.
13. *Steatohepatitis*. It is a binary variable that is equal to 1 when the patient is affected to steatohepatitis, a fatty liver disease.
Almost the 70% of the participants result positive to this pathology.
14. *HCV*. It is the acronym of Hepatitis C Virus: the variable is equal to 1 when the liver presents hepatitis C, otherwise it is equal to 0.
More than the 3/5 of the patients are positive to HCV.
15. *HBV*. It is the acronym of Hepatitis B Virus: the variable is equal to 1 when the liver presents hepatitis B, otherwise it is equal to 0.
Almost the 4/5 of the patients are positive to HBV.
16. *Potus*. It's a latin term used in medicine to indicate if the patients abuse alcohol. This variable is equal to 1 if the individual has problems with alcoholism, 0 if not.
Like *HBV*, the 4/5 circa of the participants are classified as alcohol abuser.
17. *Bilirubin*. It is a continuous variable that indicates the level of bilirubin in the blood. Normal ranges lie between 0.3 and 1.2.

18. *Albuminemia*. It is a continuous variable that indicates the level of the albumin, a protein synthesized in the liver, in the blood. Normal values range between 3.4 and 5.4.
19. *Creatinine*. It is a continuous variable that indicates the level of creatinine, a waste product, in the blood. Normal ranges are included between 0.6 and 1.3.
20. *Platelets*. It is a continuous variable that indicates the platelet count, which normally ranges from 150 to 450 thousands platelets per microliter of blood [56].
21. *INR*. It is an acronym of International Normalised Ratio and it is an index of clotting blood.
22. *inv_macro_port*. It refers to the macrovascular cancer invasion of the portal vein. It is equal to 1 when it is present, 0 otherwise.
A vast majority assumes the value 0.
23. *inv_macro_sovraep*. It refers to the macrovascular cancer invasion of other veins. It is equal to 1 when it is present, 0 otherwise.
The majority of instances assumes the value 0.
24. *Edmondson grading*. It is an histological classification that indicates the gravity of the patient's hepatocarcinoma and takes values between I,II,III and IV [57]. Higher the grade, more severe is the tumour.
25. *MVI*. It is an index of microvascular invasion, an important histopathologic prognostic factor for HCC [58]. The value 0 is associated with low risk, 1 with medium risk, 2 with high risk.
Note that some instances are equal to a white space: they are converted into NA values.
26. *Satellitosis*. It is a binary variable that indicates the presence of satellitosis when it assumes the value 1. Satellitosis is an abnormal group of cells that surrounds another singular cell and it is a risk factor for tumours.
27. *Capsule*. It is a binary variable that indicates the presence of a fibrous capsule, which is a common pathological feature in progressed HCC [59].
Unfortunately, 142 values are missing.
28. *R*. It is a categorical variable that concerns the presence of residual tumour after the surgical operation: the value 0 denotes the complete remission, the value 1 a microscopic residual tumour and the value 2 a macroscopic residual tumour [60].

% patients divided in respect to R



Since only a strict minority of the instances is equal to 2, we decide to unify 1 and 2, composing a more generic category. The value 1 now indicates the presence of residual tumour, that can be either microscopic or macroscopic. Note that some instances are equal to a white space: they are converted into NA values.

After this initial exploration of the dataset and some modifications to the features, we need to decide how to deal with **missing values**. Note that *eventomorte*, *RFS*, *timeOSmesi*, *timeRFSmesi* and *procedure* are all core variables: instances that have null values for one or more of them are then meaningless. In other words, their lacking information is too relevant to be ignored. Hence, these instances will be directly removed from the dataset: the sample reduces to 538 individuals.

We want to determine which input variables we should keep for the construction of the propensity score models and which we should discard. The choice depends on their relevance and on the number of missing values they have.

With this purpose, we build for each feature a univariate Cox model, looking at p-values, reported in 3.4.

feature	HR	lower 95% CI	upper 95% CI	p_values
eta_surg	1.005	0.996	1.015	0.291
sessom	1.169	0.893	1.53	0.257
ASA2	0.745	0.36	1.54	0.427
ASA5	0	0	Inf	0.993
CCI3	5.468	0.731	40.899	0.098
CCI12	11.067	1	122.474	0.05
MELD_score6	1.243	0.171	9.037	0.83
MELD_score27	5.393	0.335	86.741	0.234
Cirrosi1	1.292	1.031	1.619	0.026
child_pugh_grade1	1.251	0.997	1.57	0.053
Steatoepatite1	0.847	0.653	1.1	0.213
HCV1	1.088	0.858	1.38	0.486
HBV1	1.11	0.831	1.485	0.48
Potus.1	0.955	0.704	1.295	0.766
bill_tot	1.069	0.89	1.284	0.477
Albuminemia	0.787	0.639	0.969	0.024
Creatininemia	1.527	1.092	2.135	0.013
Piastrine	0.999	0.997	1	0.05
INR	1.001	0.575	1.744	0.997
inv_macro_port1	1.44	1.05	1.974	0.024
inv_macro_sovraep1	1.47	1.004	2.152	0.047
PROCEDURE1	0.993	0.734	1.343	0.963
PROCEDURE4	1.184	0.89	1.576	0.247
EdmondsonGrading2	1.095	0.691	1.736	0.698
EdmondsonGrading4	0.687	0.348	1.355	0.278
MV11	1.699	1.328	2.174	0
satellitosis1	1.56	1.173	2.076	0.002
capsula1	0.919	0.703	1.201	0.537
R1	1.056	0.757	1.472	0.749

Figure 3.4: Results of the univariate Cox models

With a level of the 95%, the variables that appear to be significant are: *cirrhosis*, *child pugh grade*, *albuminemia*, *creatinine*, *inv_macro_port*, *inv_macro_sovraep* and *satellitosis*. All of them will be included in the final models.

What we've decided to exclude is: *ASA*, *CCI*, *Meld Score* and *capsule*. They've been discarded for three main reasons: they present many missing values, they are not essential from a clinical point of view and the majority of their instances assumes only a restricted range of values.

Now, we must resolve the problem of missing values: they will be filled thanks to the multiple imputation algorithm.

3.3 Multiple imputation

There are different ways to deal with missing values. First of all, it's necessary to define which type of missing data we are handling. There are three principal categories [61]:

- *Missing completely at random*. The missing values have not systematic differences with the observed values.
- *Missing at random*. There are systematic differences between missing and observed values, but, thanks to the observed data, they can be explained.
- *Missing not at random*. Even given the observed data, the systematic differences between missing and observed values remain.

The latter is the most problematic typology, hard to deal with. We assume that our dataset missing values are members of the second category: it's therefore possible to deduce missing values from the rest of the observed data. With this purpose, we utilize the **multiple imputation** algorithm. Before implementing it in R through the *mice* package and its homonymous function [62], it's possible to examine the *pattern of missingness* of the considered dataframe.

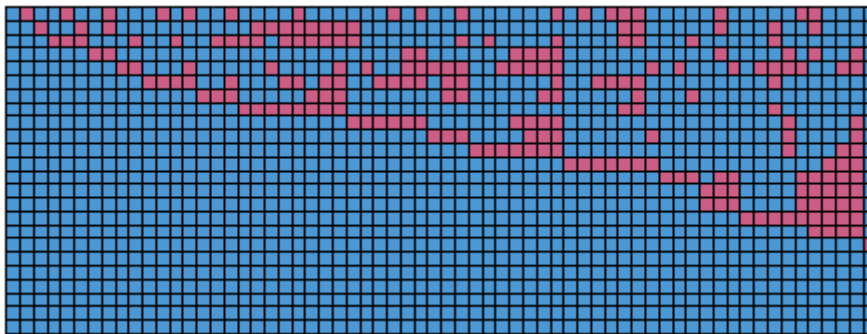


Figure 3.5: Pattern of missingness of the dataset

The grid 3.5 is composed by blue and purple squares, respectively observed values and missing values. Note that each column corresponds to one patient. Plotting this grid is recommended to verify if there are individuals with too many missing values, who are recognizable by a column full of purple squares: in that cases, the best choice is probably dropping them. Vice versa, a line composed by too many purple squares denote a variable that presents a lot of missing values.

In this case, there are some subjects with a considerable number of undefined features, but we've decided to keep all of them. In fact, the source dataset has already a limited size, so it's preferable to be more conservative.

Multiple imputation takes in input not only the features that have gaps to fill, but also the output variables and every variables linked to them, even if some of that features may be excluded from the final propensity score models. The idea is indeed to involve any data that can help in the computation of the missing values.

This method returns m datasets, that haven't got any missing data. The average of these dataframes will give us the final dataset that we are going to utilise for our study.

The univariate Cox analysis repeated with the obtained dataset shows that the significant variables are the same as before, with the addition of *platelets* and *MVI*.

Looking at the general statistics of the new dataframe, the distributions of its numerical features are almost unchanged than before the multiple imputation. By contrast, for the categorical variables, the missing values seem to be assigned to the most numerous categories. In the end, for binary features, almost all the missing values have become zeros. This makes sense, since they principally refer to pathologies and it's likely that it's more reported the presence of a disease in the patient, than its absence.

A very last modification is applied to *inv_macro_port* and *inv_macro_sovraep*: they

are unified into *inv_macro*, which indicates the presence of a macrovascular invasion of the portal vein or other veins. Therefore, it assumes the value 0 if both *inv_macro_port* and *inv_macro_sovraep* are null, 1 otherwise.

In the end, we are considering 18 descriptive variables: *age*, *sex*, *cirrhosis*, *child pugh grade*, *steatohepatitis*, *HCV*, *HBV*, *potus*, *bilirubin*, *albuminemia*, *creatinine*, *platelets*, *INR*, *inv_macro*, *Edmondson grading*, *MVI*, *satellitosis* and *R*.

3.4 Applied methods

Now, we are going to briefly describe the practical construction of the methods that we are going to utilise for the analysis of the large HCC dataset.

Note that for this case of study, we prefer estimating ATE than ATT: it is of greater interest since each surgical technique can be potentially operated to every patients. Moreover, ATE is usually put first even if the relationship between treatment and outcome is ordinal [63]. In this case, it's possible to order the categories of the resection methods from the most to the least conservative, or vice versa.

Unadjusted method

This method consists in the Cox model that regresses the time of event only on the treatment variable. We name it "unadjusted" because it is indeed not adjusted for any covariate.

It gives an estimation of the "apparent" treatment effects, i.e. how the resection techniques are associated to the overall and the relapse-free survival without considering the patients' characteristics.

Adjusted method

We call "adjusted method" the Cox model adjusted for the previously selected covariates. As we've already discussed, with observational studies, the distribution of the covariates is not homogeneous between the treatments groups, so there are confounding factors that influence the results of this model. In particular, the adjusted method estimates the *conditional* treatment effects and not the marginal treatment effects.

Matching method

The propensity score matching method ¹, described in 2.9.1, estimates only the ATT. Anyway, both ATE and ATT are measures of the marginal treatment effects.

We split the dataset in a subset containing patients that were subjected to anatomic or semi-anatomic resections and in a subset containing patients that were subjected to anatomic or wedge resections. Note that they are not disjoint sets: they have in common the individuals operated with the anatomic surgery.

In this way, we have splitted the problem into two "classical" survival analyses with

¹We sometimes use the abbreviation "cl_mn" to indicate in graphics and tables.

binary treatments. The following description illustrates the development of the model made for both the subsets.

We firstly estimate the propensity scores by a logistic regression. Then, the values estimated are used to form the matched sets.

In particular, we utilise the nearest neighbour matching, with a caliper equal to 0.3. Usually, the caliper is fixed to 0.1, but in this case of study we've decided to be more inclusive. Indeed, the HCC dataset size is already pretty small, so, we prefer to avoid a strictly selective matching algorithm not to reduce too much the dimension of the dataset composed by the matched sets.

In the end, a Cox regression is used with the obtained dataset, relating the time of the event with the treatment and also with the variable that indicates for each instance the belonging matched set.

Note that, while in the next section 3.5 we will describe the final results of the models, we are also interested in observing if the propensity score methods have actually balanced the covariates distributions. This latter analysis is then developed here.

In the graphic below, there is a comparison between the covariates distributions before and after the building of the matched sets, for the subset of patients operated with anatomic or semi-anatomic resections.

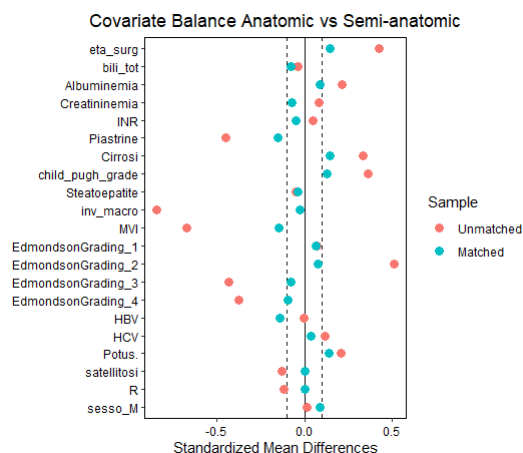


Figure 3.6: Loveplot of covariates distributions for anatomic/semi-anatomic resections balanced with propensity score matching

Only a minority of the features standardized mean differences isn't included in the range $[-0.1, 0.1]$ and the totality of them belongs to $[-0.2, 0.2]$.

An analogous result is obtained with the matched sets of individuals subjected to the anatomic or the wedge resections.

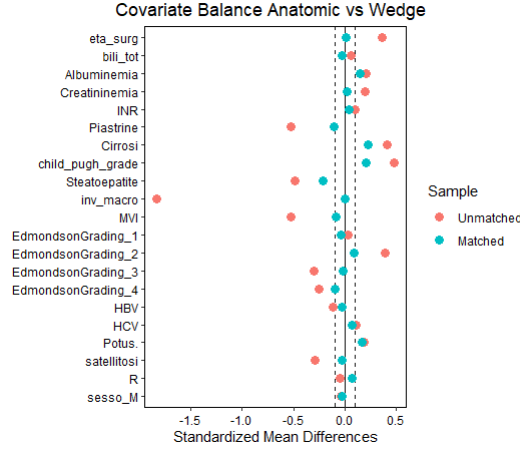


Figure 3.7: Loveplot of covariates distributions for anatomic/wedge resections balanced with propensity score matching

This time, the covariates standardized mean differences aren't contained in $[-0.2, 0.2]$, still their absolute values are never greater than 0.25 (see the figure above). We can finally ascertain that the covariates distributions have been efficiently balanced. Thanks to these graphics, we have also the clear confirmation that the covariates are not homogeneously distributed in the treatments groups and so, the adjusted method isn't appropriate to estimate the marginal treatment effects for this case of study.

IPW multinomial method

In the section 2.9.2, we described the inverse probability weighting using the propensity score ². In this case of study, we estimate the propensity score with a multinomial logistic regression and then we compute the correspondent stabilized weights to estimate the ATE. Finally, they are utilised in a weighted Cox model.

Now, let's first look at the generated weights:

²We sometimes use the abbreviation "ipw_mn" to indicate in graphics and tables.

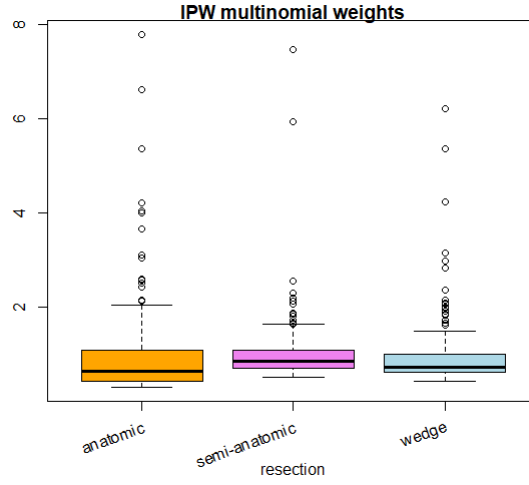


Figure 3.8: IPW multinomial weights

There is a considerable quantity of outliers, although we have utilised stabilized weights.

Then, we want to verify that the covariates have been properly balanced between the treatments groups. We first compare the covariates distribution of the patients that have been subjected to the anatomic resection with the covariates distribution of the patients who have been operated with the semi-anatomic surgery.

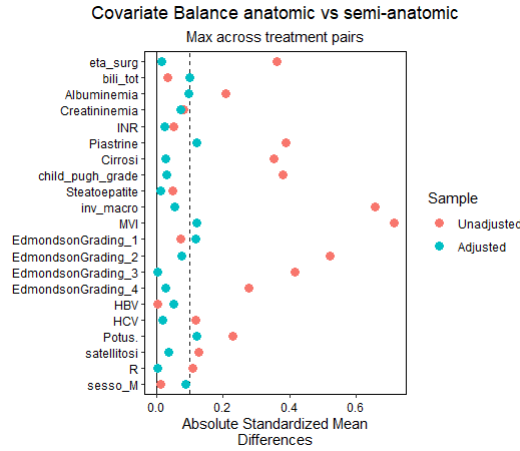


Figure 3.9: Loveplot of covariates distributions for anatomic/semi-anatomic resections balanced with IPW multinomial method

From the figure above, it is possible to see that almost every absolute standardized mean differences of the adjusted covariates don't trespass the threshold equal to 0.1. The few of them that do, still don't overcome the value of 0.2.

We also compare the group of patients that have been operated with anatomic resection

or with wedge resection.

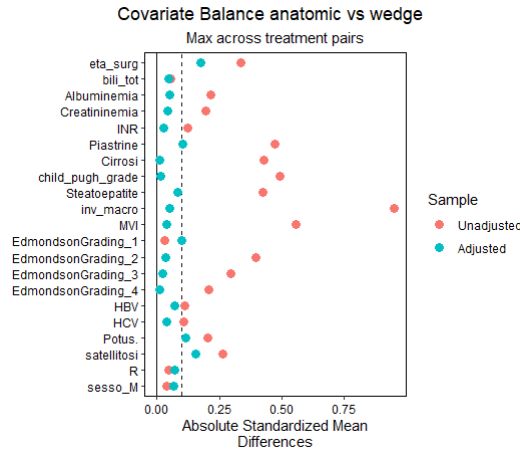


Figure 3.10: Loveplot of covariates distributions for anatomic/wedge resections balanced with IPW multinomial method

The balance observable in the figure above is worse than the previous one, anyway it still has only a strict minority of covariates characterized by absolute standardized mean differences greater than 0.1 but still lower than 0.25.

In the end, we've reached a satisfying balance.

IPW boosting method

The difference between this model and the previous one is only the fact that here we utilise the generalised boosting regression to estimate the propensity score ³. In particular, we are still estimating the ATE.

The characteristics of the weights computed thanks to this model are illustrated in the boxplot below.

³We sometimes use the abbreviation "ipw_boost" to indicate in graphics and tables.

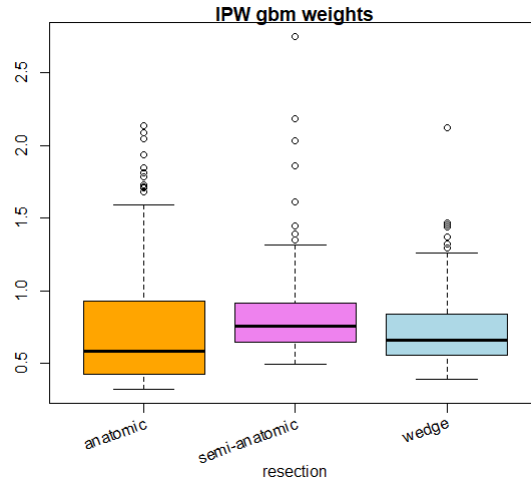


Figure 3.11: IPW boosting weights

These weights are included in a more restricted range than the ones computed with the multinomial logistic regression. Then, the probability of causing biases because of the presence of extreme weights is very low. However, the results obtained by the balancing of the covariates distributions are not so positive. Let's first focus on the comparison between anatomic and semi-anatomic resections.

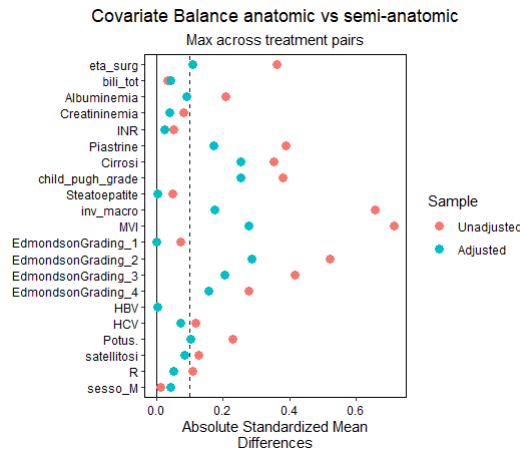


Figure 3.12: Loveplot of covariates distributions for anatomic/semi-anatomic resections balanced with IPW boosting method

Although the covariates are less unbalanced than at the beginning, there are many features with pretty high absolute standardized mean differences, as the picture above shows.

The situation results to be even worse in the comparison between anatomic and wedge

resections, illustrated in the loveplot below.

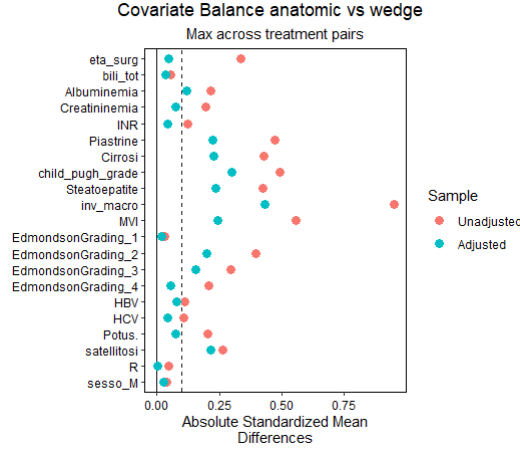


Figure 3.13: Loveplot of covariates distributions for anatomic/wedge resections balanced with IPW boosting method

At this point, we are going to put first the estimations of the treatments effects obtained by IPW multinomial rather than by this model.

GPS-CDF method

Like IPW, the generalised propensity score cumulative distribution function method ⁴ can estimate both ATE and ATT. Given the characteristics of our case of study, we prefer computing the ATE.

Following the steps illustrated in 2.9.3, we compute the generalised propensity score vector through a multinomial logistic regression, considering only the treatments order "anatomic", "semi-anatomic" and "wedge". Then, this vector is used to compute the matched sets. In this process, it is possible to choose between the optimal or the greedy matching approach. The first is more conservative than the second, which would discard some unmatched individuals. Since the study dataset has a pretty small size, we prefer the optimal matching algorithm.

Note that, while in the classical propensity score matching model the matched sets are computed separately for the two subsets previously defined, here the matched sets are computed at the same time for all the three treatments. Each individual can belong only to one matched set, which is composed by two instances. This implies that, in the end, we have three disjoint subsets: the first contains matched patients treated with anatomic or semi-anatomic resections, the second contains matched patients treated with anatomic or wedge resections and the third contains matched patients treated with semi-anatomic or wedge resections. With this construction, no reference treatment is fixed: even using the matching method, we are estimating the ATE.

⁴We sometimes use the abbreviation "gps_mn" to indicate in graphics and tables.

The construction of the Cox model is made in the same way as before with the classical matching method.

Let's finally look at the matched sets covariates distributions compared to the beginning ones.

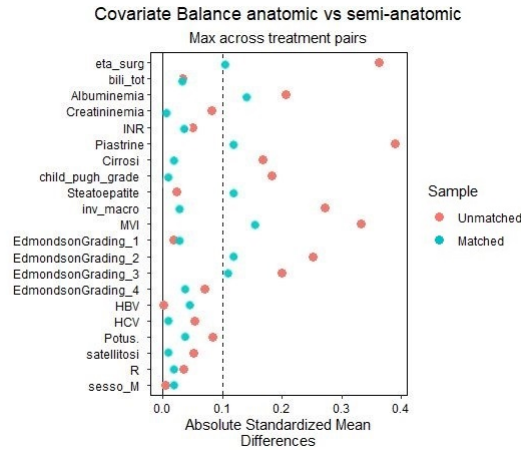


Figure 3.14: Loveplot of covariates distributions for anatomic/semi-anatomic resections balanced with GPS-CDF method

For anatomic and semi-anatomic resections, the graphic above shows a good balance of the distributions, whose absolute standardized mean differences are all included in $[0, 2]$. Unfortunately, the same can't be said for the graphic below, that concerns anatomic and wedge resections.

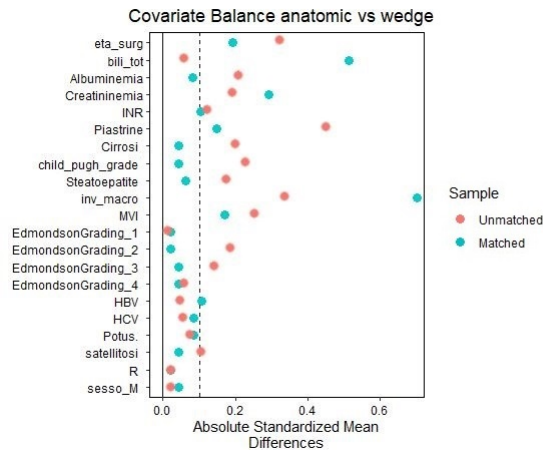


Figure 3.15: Loveplot of covariates distributions for anatomic/wedge resections balanced with GPS-CDF method

In this case, three variables are even more unbalanced than they were at the beginning. In particular, *inv_macro* has a final absolute standardized mean difference equal to 0.7. To sum up, we can observe that, overall, the number of covariates which haven't got homogeneous distributions is lower in the matched set, but there are three features that have been considerably unbalanced with the application of the GPS-CDF method. In the next chapter, for this model, we will take into account the presence of unbalance distributions in the anatomic/wedge final set.

3.5 Results

This study is born and is developed from the hypothesis that one surgical technique between anatomic, semi-anatomic and wedge is generally better than the others to cure large HCC. Our objective is to confirm or to reject this thesis and, eventually, to investigate which type of operation should be preferred. With this purpose, we first examine the results of the unadjusted Cox model for both overall and relapse-free survival.

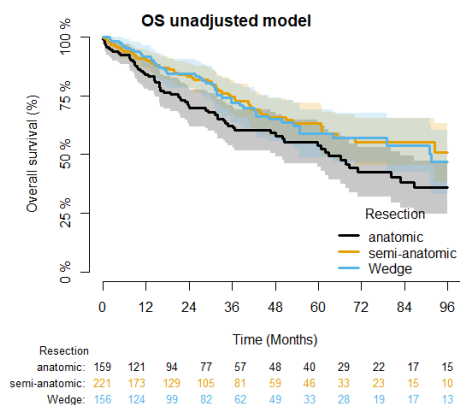


Figure 3.16: OS curve estimated with the unadjusted method

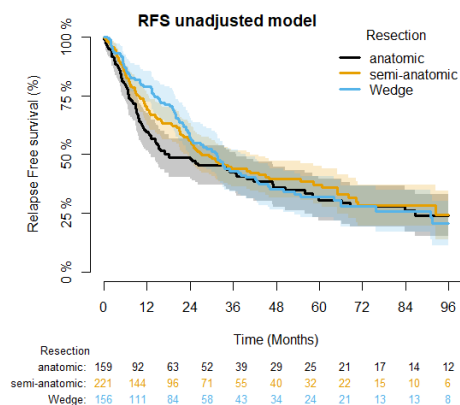


Figure 3.17: RFS curve estimated with the unadjusted method

From this first glance at the data, the techniques seem not to have different effects at all for what concerns the relapse-free survival. By contrast, the overall survival shows a significant difference between the anatomic and the semi-anatomic resections. The wedge resection, instead, isn't significantly different from the anatomic resection with a level of significance equal to 5%, as it can be seen by the p-values reported in the table 3.18 that correspond to the Wald test. Anyway, the p-value related to HR_w is still very low: it is equal to about 0.06. Moreover, the p-value associated to the score test - which, in this case, is equivalent to the log-rank test - is 0.05 (not shown in the table). Coherently with these results, the estimated \hat{HR}_s and \hat{HR}_w for the RFS-survival are included between 0.8 and 0.9, so very near to 1, while for the OS-survival are around 0.7. Hence, for the overall survival, the anatomic resection appears to be the worst surgical

technique.

Then, we look at the results of the Cox model adjusted for the previous selected variables. Note that, in the tables below, the p-values are associated to the Wald test.

Method	s_HR	s_SE	s_LOWER	s_UPPER	s_pvalue	w_HR	w_SE	w_LOWER	w_UPPER	w_pvalue
unadjusted	0,682	0,176	0,483	0,962	0,029	0,707	0,185	0,491	1,016	0,061
adjusted	0,778	0,203	0,523	1,158	0,217	0,669	0,226	0,429	1,043	0,076
ipw_mn	0,898	0,183	0,580	1,389	0,628	0,891	0,192	0,548	1,451	0,644
ipw_boost	0,749	0,203	0,511	1,098	0,139	0,802	0,216	0,537	1,197	0,280
gps_mn	0,875	0,299	0,487	1,572	0,655	0,900	0,459	0,366	2,215	0,819
cl_mn	0,885	0,286	0,505	1,550	0,668	0,522	0,356	0,260	1,049	0,068

Figure 3.18: Results of the overall survival analysis

Method	s_HR	s_SE	s_LOWER	s_UPPER	s_pvalue	w_HR	w_SE	w_LOWER	w_UPPER	w_pvalue
unadjusted	0,858	0,138	0,655	1,123	0,265	0,845	0,146	0,635	1,124	0,247
adjusted	1,074	0,156	0,791	1,459	0,647	0,979	0,173	0,697	1,373	0,900
ipw_mn	1,264	0,142	0,876	1,824	0,211	1,059	0,154	0,689	1,626	0,795
ipw_boost	1,036	0,159	0,760	1,412	0,824	0,959	0,172	0,694	1,327	0,802
gps_mn	1,118	0,236	0,704	1,775	0,638	0,571	0,362	0,281	1,161	0,122
cl_mn	1,111	0,230	0,708	1,743	0,647	0,781	0,267	0,463	1,318	0,355

Figure 3.19: Results of the relapse free survival analysis

Looking at the results of the adjusted method reported above, there is a confirmation of what we previously obtained for the RFS survival. In particular, the estimated **HR** are now ~ 1 .

On the other hand, for the OS survival, the outcomes changes in respect to before: the semi-anatomic resection completely loses the significant difference with the anatomic surgery and the \hat{HR}_s gets nearer to 0.8, while the p-value correspondents to the wedge resection compared to the anatomic one raises, but it remains low. In particular, \hat{HR}_w is estimated as 0.67, similarly as before.

Essentially, from these preliminary analyses, the choice of resection techniques seems not to be influential for the relapse-free survival, while the shape of the overall survival curve suggests the opposite for OS. Anyway, as we already discussed, these models aren't estimating the marginal hazard ratios but the conditional ones, because of the unbalanced distribution of the covariates in the treatments groups. Now, we are going to examine the results of the propensity score models, which are utilised with observational studies to mimic the randomized controlled trials and, in this way, they estimate the marginal treatment effects.

Let's start with the propensity score matching method: as already explained, it considers separately the anatomic/semi-anatomic and the anatomic/wedge subsets. First, we focus on the anatomic/semi-anatomic dataset.

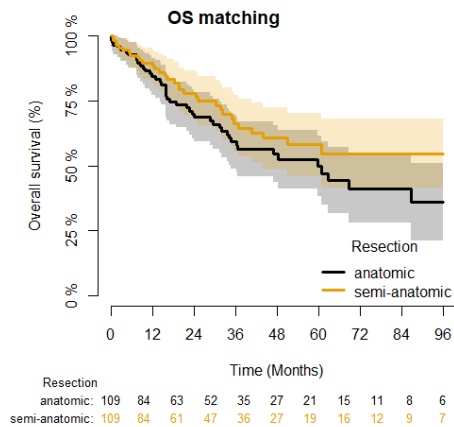


Figure 3.20: OS curve estimated with the propensity score matching method

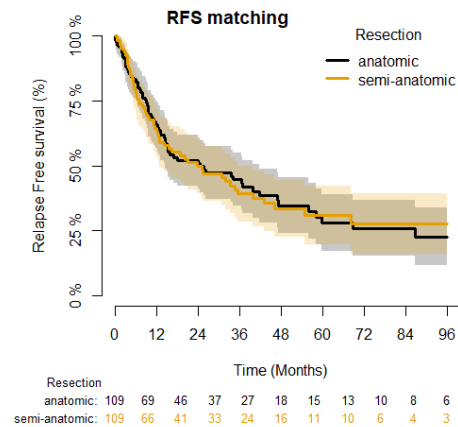


Figure 3.21: RFS curve estimated with the propensity score matching method

The relapse-free survival curves appear almost identical for the anatomic and for the semi-anatomic groups: the previous results are strongly confirmed.

The overall survival curves related to the two resections, instead, are more distant from each other than in the RFS study, but definitely rejecting the hypothesis of a significant difference between the treatments effects.

We can now observe the results for anatomic/wedge patients, illustrated in the graphics below.

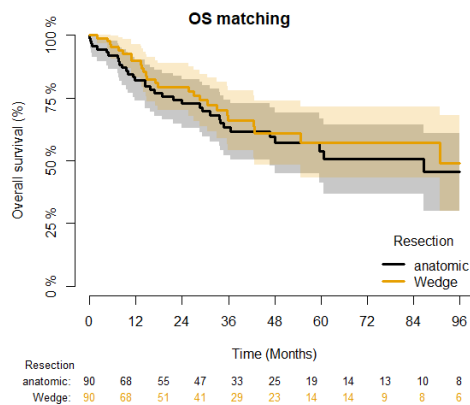


Figure 3.22: OS curve estimated with the propensity score matching method

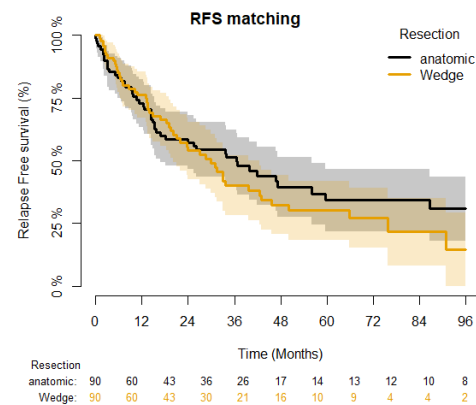


Figure 3.23: RFS curve estimated with the propensity score matching method

Here, there is still a substantial difference between the treatments groups curves for the overall survival: the correspondent p-value, even if greater than 0.05, is still low.

Note that the main flaw of this method is the loss of the sample size: the necessity of

dividing the problem into subsets implies a considerable reduction of the starting dataset dimension, which is further reduced in the final matched datasets by the matching process. Even fixing the caliper to 0.3, this problem hasn't been overcome.

The method that suffers less for the limited sample size is the inverse probability weighting using the propensity score. The numerosity of the dataset is altered by the presence of the weights and, instead of being reduced, it is usually amplified. In this context, it is a relevant positive aspect.

In the graphics below, it is possible to observe the survival curves estimated with the IPW multinomial method.

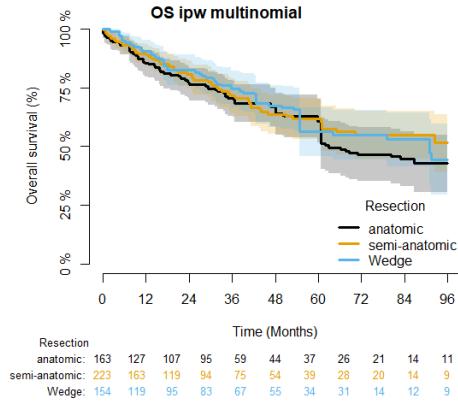


Figure 3.24: OS curve estimated with the IPW multinomial method

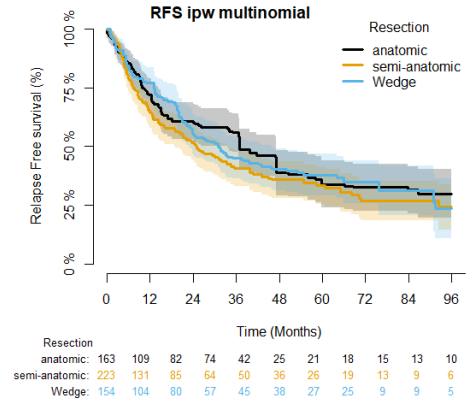


Figure 3.25: RFS curve estimated with the IPW multinomial method

In both OS and RFS, there is a stronger similarity between the semi-anatomic and the wedge resections curves than the anatomic resection curve; all of them have anyway an analogous shape. The Wald test p-values and $\hat{H}R_s$ and $\hat{H}R_w$ clearly indicate that all the estimated treatments effects aren't significantly different from each other.

Note that, now, even the anatomic/wedge overall survival curves seem not to be considerably different anymore.

The IPW boosting model shows similar results, as it can be seen in the graphics below.

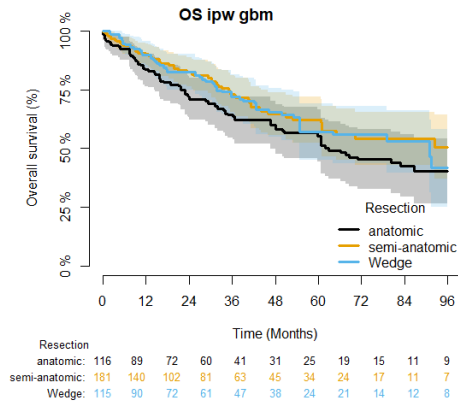


Figure 3.26: OS curve estimated with the IPW boosting method

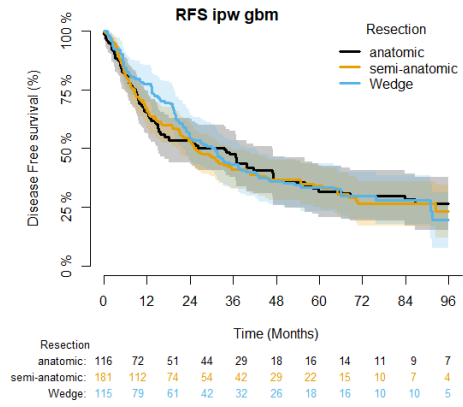


Figure 3.27: RFS curve estimated with the IPW boosting method

In particular, in the OS, the anatomic curve and the wedge and the semi-anatomic curves are less near to each other than in the previous analysis.

Analogously, as it is reported in 3.18 and in 3.19, the estimated **HR** and their p-values are lower than the ones obtained by the IPW multinomial method, but the final interpretations of the results are identical. The unique noteworthy difference is the fact that this sample has a smaller size.

Finally, we examine the results of the GPS-CDF model.

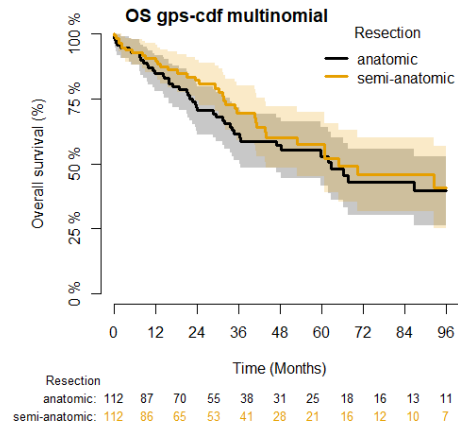


Figure 3.28: OS curve estimated with the GPS-CDF method for anatomic/semi-anatomic

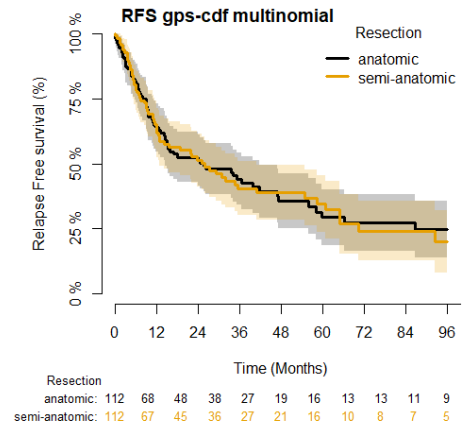


Figure 3.29: RFS curve estimated with the GPS-CDF method for anatomic/semi-anatomic

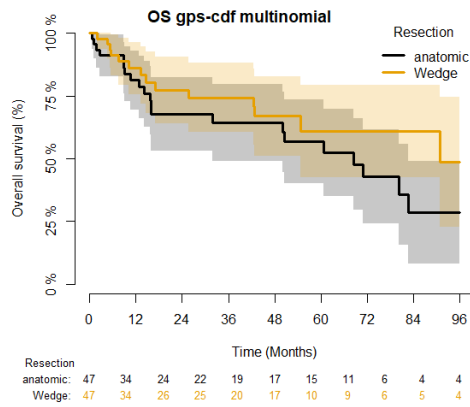


Figure 3.30: OS curve estimated with the GPS-CDF method for anatomic/wedge

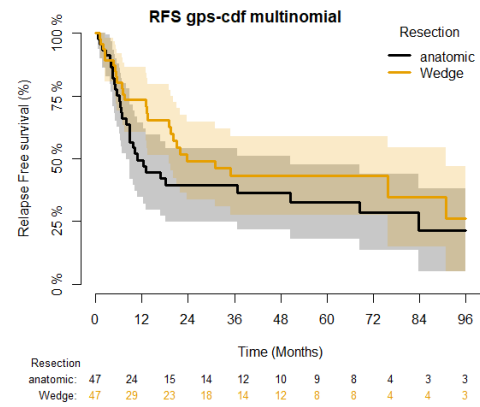
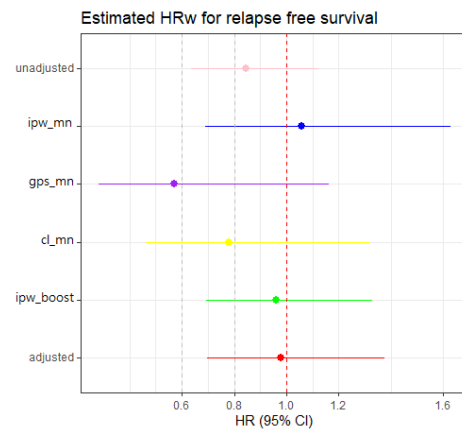
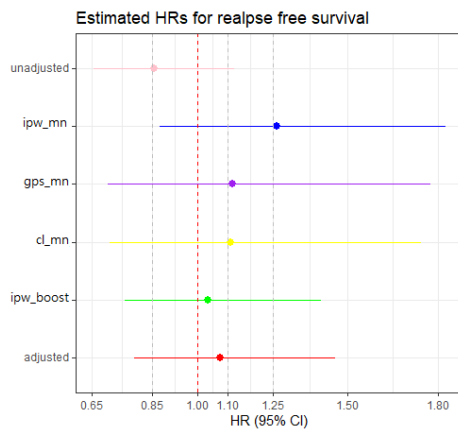


Figure 3.31: RFS curve estimated with the GPS-CDF method for anatomic/wedge

Differently from the matching method, it forms all the matched pairs - which compose the final datasets anatomic/semi-anatomic, anatomic/wedge and semi-anatomic/wedge - at the same time. We are interested only in the first two sets, whose survival curves are showed above. The hypothesis of significant differences between the treatments effects is rejected even with this model, for both OS and RFS. Anyway, what immediately stands out in the figures is the very low size of the anatomic/wedge sample. In fact, due to how the model is constructed, GPS-CDF suffers even more than the classical matching method from the problem of the reduction of the dimension of the source dataset. To conclude, we compare the estimated **HRs** and their confidence intervals obtained by each of the six models.

Let's first examine the results for the relapse-free survival, that, for every methods, doesn't show considerable differences between the treatments effects.

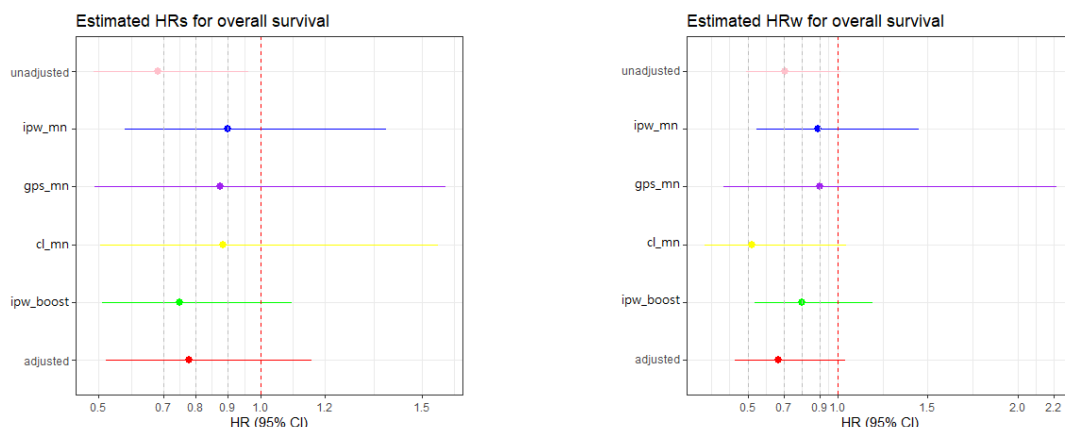


For both HR_s and HR_w , all the confidence intervals include the value 1, as it can be seen in the graphics above. This happens when the hazard ratios aren't significantly far from 1, which means that the treatments effects aren't considerably different from each other. Moreover, all the methods - except the adjusted - have $HR_s > 1$.

It's possible to notice that the GPS-CDF model estimates a particularly low value for HR_w , which could indicate a noteworthy difference between anatomic and wedge resections. Since it's the only indication of that, not even confirmed by the long-rank and the Wald tests, it is evident that it is misleading. The explanation for this anomalous result lies in the fact that the anatomic/wedge treatments group variables distributions haven't been satisfying balanced by the GPS-CDF method. Actually, some features are even more unbalanced than at the beginning, maybe also because of the very limited size of this matched group.

Finally, another aspect that emerges from the forest plots is the very large confidential intervals of IPW multinomial, GPS-CDF and matching methods.

Now, let's focus on the overall survival analysis.



The first forest plot shows a \hat{HR}_s confidence interval that doesn't include 1 only for the unadjusted method. Even the adjusted and the IPW boosting methods have pretty low values of HR_s : while the similitude between the Cox models is evident, this result is a bit anomalous for IPW. Anyway, it can be justified by the bad adjustment of the covariates distributions along the treatments groups, which in fact haven't been efficiently balanced. All the other estimations are near to 0.9 and are characterized by very large confidence intervals.

In the right graphic above, there are three methods with low values of \hat{HR}_w and whose confidence intervals include the 1 only by a little: unadjusted, adjusted and matching. In particular, the result for the matching method is far from the estimations of the other propensity score methods. Nevertheless, we saw in the previous section that the matching model reaches a good balance of the covariates distributions along the treatments groups. This time the explanation could lie in the fact that the matching method computes the ATT, instead of ATE like the other propensity score models. Then, for the comparison

between anatomic and wedge resections, the average treatment effect for the treated seems to show a considerable difference, even if not very pronounced. In particular, the \hat{HR}_w estimated by the matching method is equal to 0.5.

3.6 Discussion

The hepatocellular carcinoma represents a worldwide health challenge, due to its high mortality and relapses rate. One of the principal utilised therapies is the surgery, whose study is the core of this thesis. Given a dataset of patients affected by large HCC, we use causal inference methods with the purpose of determining if one of the resection techniques can be considered better than the others. In particular, we classified the surgical operations in three categories: anatomic, semi-anatomic and wedge resections. We set up both an overall and a relapse free survival analyses and presented the results above. The unadjusted and the adjusted models estimate the *conditional* treatment effect, i.e. what we defined in 2.8.4 as the average effect of subjecting a patient to a specified treatment. It is referred to the individual level and not to the population one. What we actual aim to is the *marginal* treatment effect, because it gives a clearer indication of how the chosen therapies influence the final outcomes. In fact, even if the conditional effect shouldn't be considered as meaningless in itself, it is much harder to be correctly interpreted.

In the unadjusted Cox model, the outcomes are only regressed on the treatment variable: we are looking for differences between the times of event related to each resection techniques in a completely general frame. By contrast, the adjusted Cox model takes into account also the covariates classified as influent.

From the obtained results, we can deduce that the conditional treatments effects aren't considerably different in respect to the resection techniques for the relapse-free survival. The overall survival, instead, shows a difference between the anatomic and the wedge resections which is near to the significance threshold and so can't be ignored. For what concerns anatomic and semi-anatomic surgeries, the Wald and the long rank tests null hypotheses are rejected only by the unadjusted method, the weakest and least reliable. Once also the patients' covariates are taking into account in the model, $H_0 : \beta = 0$ isn't refused anymore.

The propensity score methods are then utilised to estimate the marginal treatment effects. They are theoretically more suitable to answer our research goal. Unfortunately, they share a relevant issue: their efficiency is reduced because of the limited size of the study sample. Underlined this difficulty, the IPW multinomial method stands out compared to the others, since it keeps the source dataset dimension - actually, thanks to the weights, its dataset size is even greater than the original -. Moreover, it reaches a good balance in the covariates distributions along the treatments groups. Finally, its construction is also relatively simple, with a restrained computational cost. It works even better than the IPW boosting model, not too brilliant in this case, which suggests that the relationship between the covariates and the treatments is linear. Its unique noteworthy flaw is the large range of its confidence intervals, which is nevertheless practically irrelevant

this time, as it can be deduced soon.

The IPW multinomial model estimates the ATE, considered a measure of the marginal treatment effect, and all its results show that there is no significant difference between the treatments effects, for both overall and relapse-free survival. Our final conclusion embraces these findings: no resection techniques can be considered as generally better than the others to cure large HCC.

In light of this, the results obtained by the Cox models for the overall survival analysis are probably caused by the presence of confounding factors that influence the outcomes. For instance, one explanation could be that patients with worse health conditions are more likely to be subjected by an anatomic than a wedge resection.

The propensity score method which instead seems to be the worst is the GPS-CDF: it is in fact the most penalized by the small sample size. Since it generates three distinguished matched datasets from the same total source dataset, the size of the final samples are potentially smaller than the ones obtained by the classical matching algorithm. Unfortunately, this weak point largely compromises the GPS-CDF well functioning in the comparison of anatomic and wedge resections, since the computed adjustment of their groups covariates distributions turns out to be inefficient. Anyway, this is not a sufficient reason to discard completely this method, since it has interesting theoretical basis for survival analysis in multi-treatment settings, more solid for some aspects than the IPW and the matching models. In further studies with bigger datasets, its contribution could be relevant.

About the propensity score matching, we have a result that can't be ignored for the estimation of HR_s in the overall survival analysis. The null hypothesis of the Wald test is not rejected, but the correspondent p-value is low and $\hat{H}R_w$ is equal to 0.5, facts that suggest the presence of a considerable difference between the anatomic and the wedge treatment effects. Moreover, this time this result can't be attributed to an anomalous functioning of the method, since there is a good balance between the covariates distribution along the patients' groups.

This divergence from the other propensity score results can be explained by the fact that matching method estimates ATT and not ATE, even if they are both considered estimations of the marginal treatment effect. Having found a low value of ATT indicates that if we had operated patients, who actually were subjected to the wedge resection, with the anatomic surgery, the effects would have been worse in terms of overall survival. This result, added to the ones obtained from the unadjusted and the adjusted Cox models, are indicative of the importance of the choice of the resection techniques to use for the overall survival. Even if there isn't a technique generally better than the others with large HCC, it is reasonable that it's possible to define the best surgical method to operate on the basis of the patients' characteristics, besides the tumour dimension. This conclusion is supported also by a clinical point of view, since these resection techniques shouldn't be equivalent to each other. Hence, in order to investigate this aspect, a prediction analysis and a classification model should be developed: the first, given a set of covariates, aims to predict the overall survival time, while the second gives in outcome the best resection technique to use for the considered individual. Anyway, these analyses aren't trivial and

they would encounter the issue of the limited sample size. Then, instead of considering only a subset of the data collected on the HCC Italian Register, further studies should probably include almost the entire dataset, if possible.

The fact of having taken into account only a set of patients who suffer from large hepatocellular carcinomas can also be the reason why the RFS analysis resulted to be inconclusive. Indeed, it is possible that all the tumours included in the research are in an advanced stage such that they very often request other therapies to get to the healing. Then, the relapse-free survival analysis should be reconsidered with larger datasets, evaluating also the possibility of redefining the concept of relapsing. An example is the introduction of a level of the recurrence severity.

Now, let's observe the performances of the models applied to simulated datasets.

4. Simulation study

4.1 Monte Carlo method

After having applied propensity score methods to our case of study, we need to test the quality of the utilised methods, firstly to evaluate the reliability of the models, then to compare their performances. With this purpose, we build a simulation study, generating 1000 datasets, with 1000 instances each.

We utilise the **Monte Carlo method**, which was developed at the end of the II World War, thanks to an intuition of the physicist Stanisław Ulam. He elaborated the idea while he was working in a research group that was studying a computational model of a thermonuclear reaction, as it is reported by Nicholas Metropolis in [64]. In particular, this method was described for the first time in 1949 by Ulam and Metropolis, in the paper [65].

The main idea of Monte Carlo is to generate a multiple random sampling from a given set of probability distributions, stabilizing also the relationship between the input variables. Therefore, we need to define which variables are taken in input, their distribution functions, their interactions and their coefficients.

In this case we are considering a categorical output, i.e. the treatments variable. Hence, given the input features, we compute for every instances three probabilities: the probability of being operated respectively with the anatomic, the semi-anatomic or the wedge resection. Finally, a sample function returns the outcomes, using these probabilities as parameters.

4.2 Study design

For the Monte Carlo method, it would have been problematic considering all the 18 descriptive features used before, because of the computational costs and the difficulty of simulating so many variables. For this reason, we selected the 10 of them that we valued as the most important: *age*, *sex*, *cirrhosis*, *steatohepatitis*, *HCV*, *HBV*, *creatinine*, *inv_macro*, *MVI* and *satellitosis*. Note that all the variables that resulted significant during the univariate analyses hadn't been discarded.

To generate them, we choose to fix their distributions as all binomial, except for the continuous variables - age and creatininemia -, which have been associated to a normal distribution. For the latter, the parameters inserted into the R function (we utilised *rnorm* [66]) are the mean and the deviation standard of the features, computed for the large HCC dataset 3.2. By contrast, the categorical features are reproduced by the R function *rbinom* [67], that takes as input their probability of assuming the value 1, still computed for the previous dataset of study.

The probabilities of the three possible outcomes (the anatomic, the semi-anatomic and

the wedge resection) are calculated by the inverse logit function [68]:

$$\text{logit}^{-1}(x) = \frac{e^x}{1 + e^x}$$

Note that it maps $x \in \mathbb{R} \rightarrow [0, 1]$.

With binary settings, it's sufficient to utilise this function to compute one of the treatments probabilities, while the other is just its complementary. Conversely, with multiple treatments, it is necessary to compute $n_t - 1$ probabilities. In our case, we decided to estimate the semi-anatomic and the wedge resections probabilities, respectively p_s and p_w , with the inverse logit function and, then, to compute the anatomic surgery probability as $p_a = 1 - p_w - p_s$. The problem is that the result ranges between -1 and 1, so we put the condition that, if $p_w + p_s \geq 1$, p_a is equal to 0. Here there is the input argument to the inverse logit function:

$$x = \alpha_{0,t} + \alpha_{a,t} \cdot \text{age} + \alpha_{c,t} \cdot \text{creat} + \alpha_{s,t} \cdot \text{sessom} + \alpha_{ci,t} \cdot \text{cirrhosi} + \alpha_{st,t} \cdot \text{steato} + \\ \alpha_{hb,t} \cdot \text{HBV} + \alpha_{hc,t} \cdot \text{HCV} + \alpha_{sa,t} \cdot \text{satellitosis} + \alpha_{i,t} \cdot \text{inv} + \alpha_{m,t} \cdot \text{MVI}$$

where t is equal to the semi-anatomic or to the wedge resection. The input variables (age, creat, etc.) are the ones generated as described above. The coefficients α are, instead, adjusted to generate input and output variables with characteristics as more similar as possible to the original ones.

Remember that p_a , p_s and p_w are given as inputs of a sample function and in the end we will obtain the treatment vector, composed by 1000 categorical values that belong to $\{\text{anatomic}; \text{semi-anatomic}; \text{wedge}\}$.

The last step of the simulation is generating the time to event outcomes. It's necessary to fix a distribution for them: we've decided to use a **Weibull** function, calibrating its two parameters h and p to obtain outputs that fit the original ones. In particular, the chosen values have been $h = 0.0035$ and $p = 1.5$. In this way, the time to event for the semi-anatomic treatment is computed with the inverse formula of the survival Weibull:

$$t.event.s = -\frac{\log(U)^{\frac{1}{p}}}{h \exp(Q_s)}$$

where $U \sim \mathcal{U}(0, 1)$ and

$$Q_s = \beta_{0,s} + \beta_a \cdot \text{age} + \beta_c \cdot \text{creat} + \beta_s \cdot \text{sessom} + \beta_{ci} \cdot \text{cirrhosi} + \beta_{st} \cdot \text{steato} + \beta_{hb} \cdot \text{HBV} + \\ + \beta_{hc} \cdot \text{HCV} + \beta_{sa} \cdot \text{satellitosis} + \beta_i \cdot \text{inv} + \beta_m \cdot \text{MVI}$$

The generation of the time of event for the wedge treatment is analogous.

The coefficients β are inspired by the coefficients of the multivariate Cox model with the same input variables and $(t.event, event)$ as outcome (where the event is relapse or death). While the intercept for the anatomic treatment is null, $\beta_{0,s}$ is equal to the logarithm of the *conditional* hazard ratio for the semi-anatomic resection. The problem is that our propensity score models aim to estimate the *marginal* hazard ratio. Therefore, we are interesting in building datasets with specified marginal hazard ratios, that

we call HR_s and HR_w and are respectively equal to $\frac{\lambda_{\text{semi-anatomic}}}{\lambda_{\text{anatomic}}}$ and $\frac{\lambda_{\text{wedge}}}{\lambda_{\text{anatomic}}}$. With this purpose, we utilise two iterative processes, one for HR_s and the other for HR_w , that calibrate the intercepts $\beta_{0,s}$ and $\beta_{0,w}$ in such a way that their induced marginal hazard ratios \widetilde{HR}_s and \widetilde{HR}_w will almost equal to the fixed HR_s and HR_w (the same idea used in [18]). In particular, the cycle stops when $|\widetilde{HR}_t - HR_t| < 10^{-3}$ or when it reaches a specified maximum number of iterations. Whenever the maximum number of iterations is reached, the generated dataset is considered not suitable for the chosen hazard ratios and, then, discarded. This happens to avoid too long loops.

The starting values of $\beta_{0,s}$ and $\beta_{0,w}$ have been arbitrarily put equal to $\log(HR_s)$ and $\log(HR_w)$ before entering the iteration process.

Finally, even the censoring is simulated: it must be non-informative, as observed in 2.5, consequently we generate it as an uniformly distributed variable. Since only a small percentage of the instances should be subjected to it, we fixed as minimum value 96 months (which corresponds to 8 years).

So far, we completed the description of the datasets generation process. We are going to develop 4 simulations: case A with $(HR_s = 0.8, HR_w = 0.5)$, case B with $(HR_s = 1.5, HR_w = 0.5)$, case C with $(HR_s = 1.5, HR_w = 2)$ and case D with $(HR_s = 0.2, HR_w = 0.7)$ (see 4.1). For all of them, the percentages of censoring is very low.

We are going to apply the propensity score methods described above to every cases. In particular, to analyse the methods performances, for each case we consider the mean estimated hazard ratios, their confidence intervals, the coverage (i.e. the percentages of the C.I. that contain the real value of HR), the mean standard errors of the log hazard ratios across the 1000 simulated datasets and the empirical standard error, i.e. the standard deviation of the estimated log hazard ratios computed across the 1000 simulated datasets.

The code of the simulation study can be found here [69].

case	HRs	HRw
A	0.8	0.5
B	1.5	0.5
C	1.5	2.0
D	0.2	0.7

Figure 4.1: Table with the marginal hazard ratios fixed for each simulated case

4.3 Results

Let's make a summary of the quantities we are going to estimate and compare to analyze the algorithms performances:

1. **Mean of HR vs marginal treatment effect.** For each simulated dataset, every model estimates two hazard ratios - one for anatomic in respect to semi-anatomic resection and the other for anatomic in respect to wedge resection -, so 1000 \widetilde{HR}_s

and 1000 $H\hat{R}_w$. Then, we consider their mean and compare it to the real HR_s and HR_w . In this way, we want to verify the reliability and the precision of the estimations made by each method.

2. **Confidence interval for the HR estimation and its length.** For every methods, we compute the 95% confidence intervals that correspond to each estimation of the hazard ratio. In the end, we consider the mean of these intervals, i.e. a range that has as lower extreme the mean of the computed C.I. lower extremes and as upper extreme the mean of the computed C.I. upper extremes. Since we know the real values of HR_s and HR_w , we control if these average confidence intervals actually contain them.

Note that, if methods work well, we have the 95% of confidence that $HR_s \in \overline{CI}_s$ or $HR_w \in \overline{CI}_w$. Then, the length of these intervals is an indication of our estimations uncertainties. Therefore, large C.I. are more likely to contain the real values of the marginal hazard ratios, but at the same time give an estimation of the **HR** too vague. On the contrary, short intervals are more precise and indicate a limited reference range for **HR**, but the risk of missing HR_s and HR_w is higher.

3. **Coverage.** As we already said, theoretically, the confidence intervals computed should contain the real values of the marginal hazard ratios with a level of confidence of the 95%. In other words, in each simulation, HR_s and HR_w should be included in the 95% circa of the C.I. estimated for every methods. For this reason, we verify the coverage for each case, i.e. the percentage of the ranges that actually contain the marginal **HR**.

Results that are particularly lower than 95% indicate relevant inaccuracy and presence of issues in the considered method.

4. **Mean of standard errors of the $\log(\mathbf{HR})$ vs empirical standard error.**

It is desirable that the first quantity estimates the second one, which indicates the sampling variability of the estimated $\log(\mathbf{HR})$. Therefore, we will consider their ratio: if it is near to 1, the average of the standard errors of the $\log(\mathbf{HR})$ is estimating correctly the empirical standard error, while a result greater than 1 corresponds to an overestimation, lower than 1 to an underestimation.

There is a hierarchy in the importance of these quantities with the purpose of evaluating the performances of the methods: for example, the last one is the least relevant. Anyway, each of them is interesting and only after observing them as a whole, it is possible, given a specific context, to identify the best method.

We utilise 5 of the models built in the analysis of the HCC dataset: the *adjusted model*, the *inverse probability weighting using the propensity score estimated by a multinomial logistic regression*, the *inverse probability weighting using the propensity score estimated by a generalised boosted regression*, the *generalised propensity score cumulative distribution function method* and the *propensity score matching method*. In this case, it's useless to consider the unadjusted method.

Our goal is to estimate the marginal treatment effects. Before examining the results we

are going to obtain, some preliminary observations can be already made. First of all, from its proper definition, the Cox model adjusted by the covariates isn't appropriate to estimate marginal effects, but conditional ones. In fact, it measures the effect of a treatment given the patients' characteristics and, in particular, treated population and untreated population are profoundly different. Therefore, we already know that this model is *not* estimating the marginal hazard ratio. For this reason, it is considered as the *model zero*, the simplest model that in these scenarios should work worse than the others.

Another important note that can be highlighted before the analysis is the fact that the propensity score matching model estimates ATT, while the other three propensity score methods estimate ATE. Which one between ATE and ATT is an estimation of the marginal treatment effect? The answer is *both*. They have different nuances of meaning and what is better to consider between them depends on the clinical context, but it's not wrong to compare them as measures of the marginal treatment effect. Anyway, in the examination of the results, it is advisable to take this difference into account.

Finally, before the analyses, we shortly describe the models we are going to utilise, that have some readjustments in respect to the ones used for the large HCC data.

The IPW models don't change: they use stabilized weights and robust Cox regressions; no trimming threshold has been fixed. It has indeed been noticed that only a strict minority of extreme weights is generated.

For the GPS-CDF method, we utilise the greedy matching function, instead of the optimal matching function used in the study case. Analogously, the matching models have a caliper equal to 0.1 - the standard one - and not equal to 0.3 as before. In fact, we haven't problems with the sample size anymore, so we prefer to be more selective.

Now, let's look at the models results.

Estimated marginal hazard ratios and their confidence intervals

We are considering four cases, trying to combine different values of marginal hazard ratios to cover a large variety of scenarios. In two of them, the semi-anatomic resection has worse effects than the anatomic one (when **HR** > 1), while, in the other two, it is true the opposite (when **HR** < 1).

In particular, in the case A, the hazard ratio is near to 1, so there is a similarity in the treatments effects. This last case is the one where the adjusted model estimates best the marginal HR_s (see 4.2): this is reasonable, because the conditional effect is likely to be nearer from the marginal effect when the distance between the effects of the compared treatments is smaller.

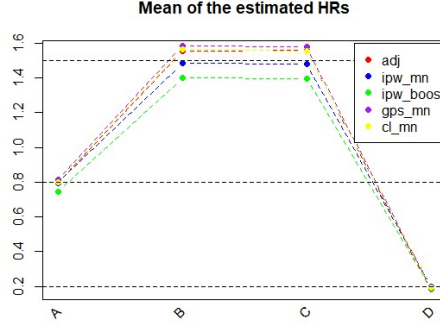


Figure 4.2: Estimations of HR_s compared in the four cases

Coherently with the reasoning above, the case where $HR_w = 2$ is the worst case for the adjusted model estimation of the marginal HR_w , as it is shown in 4.3. Indeed, this value of the hazard ratio indicates that the anatomic resection has effects largely better than the wedge resection.

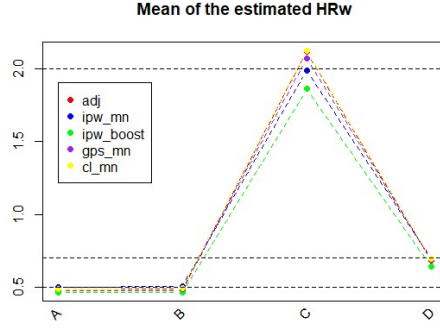


Figure 4.3: Estimations of HR_w compared in the four cases

Moreover, in case D, the marginal HR_s has a very low value, which is an indication that the semi-anatomic resection works much better than the anatomic resection. In this last case, the real value of HR_s isn't even contained by the interquartile range of the hazard ratios estimated by the adjusted model, as it can be seen in 4.4.

Considering the overall results, the IPW that uses the generalised boosted regression is definitely the worst model. It is clearly not adapt to analyse the cases considered. In particular, this method always underestimates the marginal hazard ratio. That shouldn't alarm us too much: as we said in 2.10.2, the gbm regression works well with non-linear treatment models, while it doesn't fit the simpler cases and our simulations assume a linear relationship between the treatments and the covariates.

In 4.4 and 4.5, we can find also the confidence intervals of the estimated hazard ratios: as we expected, the shortest ones belong to the adjusted models. Anyway, even the IPW methods have pretty short confidence intervals.

Unfortunately, the other two methods are denoted by very large confidence intervals:

in particular, GPS-CDF models are characterised by the widest ones and by the higher number of outliers.

Note that there are cases where, while the values of HR_s remain unvaried, the values of HR_w change and vice versa. As it can be observed in the figures below, in these cases the propensity score methods don't estimate differently the same **HR**, independently from the other hazard ratio considered.

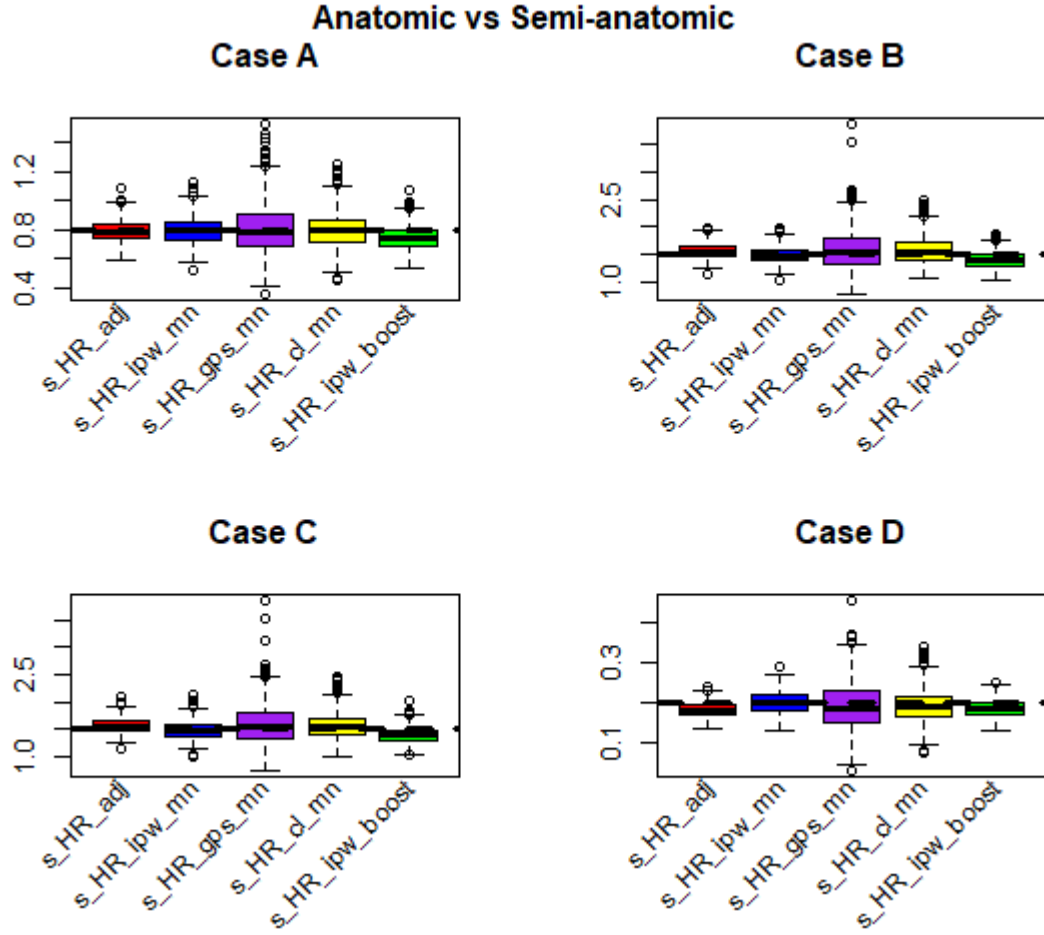


Figure 4.4: Estimations of HR_s made with 5 different methods

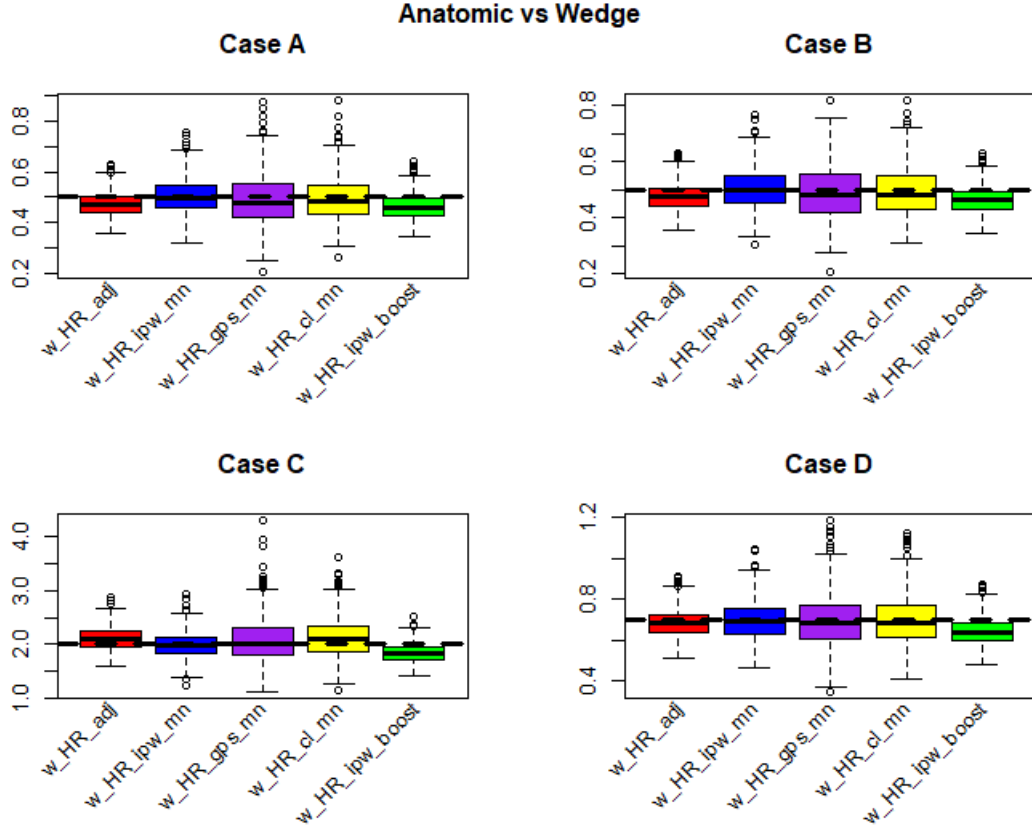


Figure 4.5: Estimations of HR_w made with 5 different methods

Coverage

Since we are considering 95% confidence intervals, we expect from our models a coverage near to the 95%. In other words, our will is to find that at least the 95% of the estimated confidence intervals contains the real value of the marginal hazard ratio.

These expectations are completely unsatisfied by the IPW boosting model, that we already pointed out as not adapted to these problem cases: it doesn't even reach the 90% of correct confidence intervals, as it can be seen in the figures 4.6,4.7.

Also the adjusted model has never been able to achieve the threshold of the 95%, even if it is often near to it. In case D, when $HR_s = 0.2$, it is the model with the lowest percentage of coverage. As we already observed, this happens probably because the conditional treatment effect is particularly far from the marginal one.

Remember that the adjusted model confidence intervals are the shortest, so their risk of missing the real marginal hazard ratios is higher than the other models.

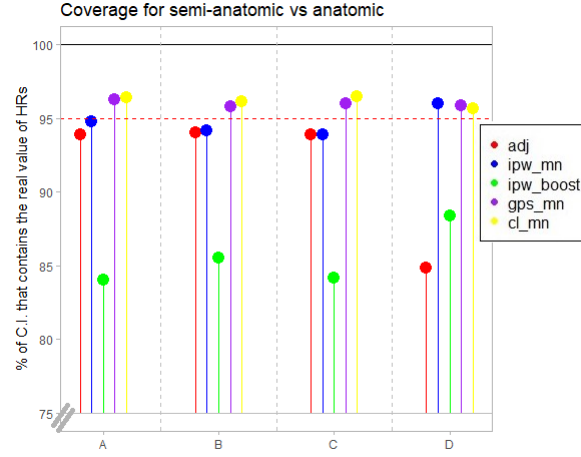


Figure 4.6: Coverage for HR_s with 5 different methods

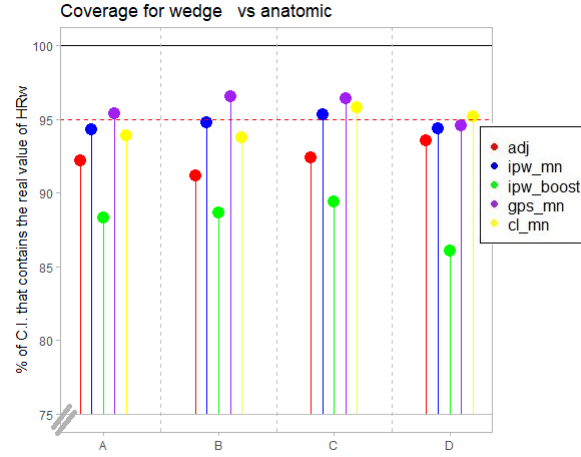


Figure 4.7: Coverage for HR_w with 5 different methods

The GPD-CDF method has the highest percentages of coverage: this is coherent with the fact that it is characterized by the largest confidence intervals. Anyway its coverage values are analogous to the ones obtained with IPW multinomial and matching models. That is a relevant point in favour to IPW multinomial method: it has shorter C.I., good estimations of \mathbf{HR} and, still, satisfying percentages of coverage.

Estimated sampling variability of $\log(\mathbf{HR})$

The sampling variability of the estimated $\log(\mathbf{HR})$ is given by the mean of the standard errors computed for all the 1000 simulated datasets. We expect that the standard deviation of the 1000 logarithms of the estimated hazard ratios corresponds to this variability.

For this reason, we compute the ratio $\frac{\text{mean}(se)}{\text{sd}(\log(\mathbf{HR}))}$, willing that it results to be near to 1. Only two methods show a deficit in this estimation: IPW multinomial and IPW boosting, as it can be noted in the graphics below.

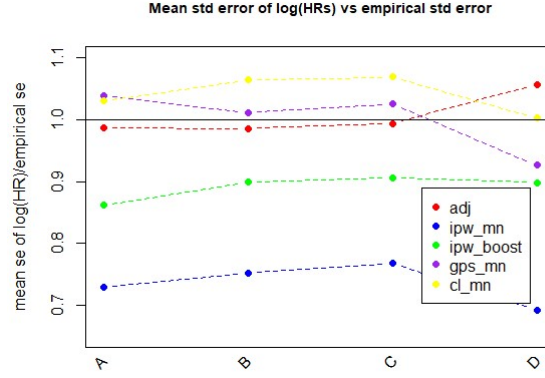


Figure 4.8: Ratio between the standard deviation of $\log(\hat{H}\hat{R}_s)$ and the empirical standard error

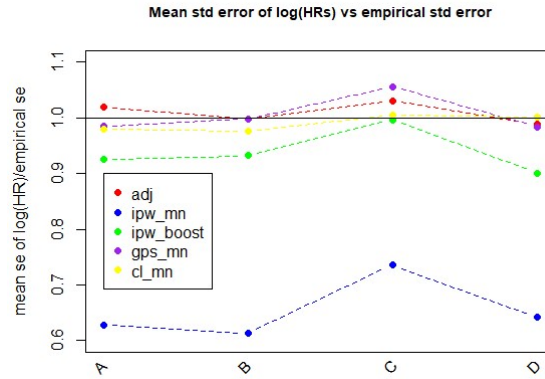


Figure 4.9: Ratio between the standard deviation of $\log(\hat{H}\hat{R}_w)$ and the empirical standard error

Both of them tend to underestimate the sampling variability of $\log(\mathbf{HR})$, but the lowest ratios computed belongs to IPW multinomial. In particular, there is a considerable distance between the ratios obtained from the IPW multinomial model and all the other ratios (see 4.8,4.9).

Note that the adjusted method, which has been showed to work poorly, seems to perform well for this estimation. This is a clear signal that we can't rely on this performance measure to find the best models: methods that aren't satisfying for the previous criteria should be discarded, independently from the results obtained for this estimation.

Anyway, showing good estimations of the sampling variability of $\log(\mathbf{HR})$ is a point of favour for the methods that have already been proved to perform well.

Overall results

To identify the best methods, it is necessary to consider all the previous performance measures as a whole. Moreover, they can assume different levels of importance on the basis of the clinical context and the research goal.

Since the beginning, we knew that the adjusted model isn't adapt for the estimation of the marginal treatment effects of observational studies, because it doesn't balance the covariate distributions in the treatment groups. It has been included in the analysis to take the role of the "*model zero*", i.e. a reference model that should work worse than the others.

Actually, the IPW boosting method has been showed to be the worst, having really bad performances for every measures except the estimations of the sampling variability of $\log(\mathbf{HR})$. We can then conclude that the generalised boosted regression doesn't work for these simulation cases, probably because the generated relationships between treatments and covariates are too simple for gbm.

Moreover, we haven't already mentioned the computational costs, but the IPW boosting is largely more expensive than the others.

Even GPS-CDF requests a consistent execution time: it is indeed the one with the most complicated construction, since it considers the whole generalised propensity score vector. Anyway, its performances aren't particularly efficient: its main flaw is the confidence intervals length.

In the end, except for its estimations of the sampling variability of the $\log(\mathbf{HR})$, the model that seems to work as the best is IPW multinomial. In fact, it makes almost unbiased estimations of HR_s and HR_w with pretty strict confidence intervals, still reaching acceptable percentages of coverage. Even matching method isn't characterized by bad results, but its confidence intervals are unfortunately too large.

The tables 4.10, 4.11, 4.12 and 4.13 contain the overall numerical results for each simulation case.

Method	HR_semi	lower_semi	upper_semi	dev_semi	se_semi	ratio_semi	s_coverage (%)	HR_wedge	lower_wedge	upper_wedge	dev_wedge	se_wedge	ratio_wedge	w_coverage (%)
adjusted														
ipw_mn	0.793	0.662	0.950	0.093	0.092	0.988	93,900	0.475	0.392	0.576	0.096	0.098	1.020	92,200
ipw_boost	0.794	0.638	0.989	0.110	0.080	0.728	94,800	0.503	0.384	0.661	0.138	0.087	0.629	94,300
gps_mn	0.744	0.609	0.910	0.102	0.088	0.862	84,000	0.464	0.373	0.578	0.105	0.097	0.926	88,300
gps_mn	0.812	0.521	1.265	0.218	0.227	1.038	96,300	0.496	0.335	0.707	0.195	0.192	0.984	95,400
cl_mn	0.798	0.591	1.078	0.149	0.154	1.020	96,400	0.492	0.353	0.685	0.173	0.170	0.980	93,900

Figure 4.10:
Case A

Method	HR_semi	lower_semi	upper_semi	dev_semi	se_semi	ratio_semi	s_coverage (%)	HR_wedge	lower_wedge	upper_wedge	dev_wedge	se_wedge	ratio_wedge	w_coverage (%)
adjusted														
ipw_mn	1.554	1.296	1.864	0.094	0.093	0.986	94,070	0.474	0.390	0.576	0.099	0.099	0.998	91,156
ipw_boost	1.483	1.200	1.836	0.107	0.080	0.751	94,171	0.504	0.383	0.665	0.143	0.088	0.613	94,774
gps_mn	1.399	1.151	1.701	0.098	0.088	0.898	85,238	0.464	0.373	0.578	0.106	0.099	0.932	88,643
cl_mn	1.584	1.005	2.499	0.228	0.231	1.011	95,779	0.487	0.335	0.707	0.192	0.192	0.998	96,583
	1.563	1.150	2.124	0.147	0.156	1.064	96,181	0.492	0.354	0.685	0.174	0.170	0.976	93,769

Figure 4.11:
Case B

Method	HR_semi	lower_semi	upper_semi	dev_semi	se_semi	ratio_semi	s_coverage (%)	HR_wedge	lower_wedge	upper_wedge	dev_wedge	se_wedge	ratio_wedge	w_coverage (%)
adjusted														
ipw_mn	0.181	0.147	0.222	0.099	0.105	1.057	84,855	0.683	0.565	0.825	0.097	0.096	0.988	93,581
ipw_boost	0.200	0.153	0.263	0.133	0.092	0.691	95,988	0.695	0.536	0.902	0.132	0.085	0.643	94,383
gps_mn	0.185	0.147	0.234	0.114	0.103	0.898	88,365	0.644	0.519	0.799	0.105	0.095	0.901	86,058
cl_mn	0.190	0.105	0.346	0.341	0.316	0.926	95,888	0.691	0.483	0.987	0.186	0.183	0.983	94,584
	0.191	0.127	0.286	0.210	0.210	1.003	95,687	0.695	0.506	0.954	0.162	0.162	1.001	95,186

Figure 4.12:
Case C

Method	HR_semi	lower_semi	upper_semi	dev_semi	se_semi	ratio_semi	s_coverage (%)	HR_wedge	lower_wedge	upper_wedge	dev_wedge	se_wedge	ratio_wedge	w_coverage (%)
adjusted														
ipw_mn	0.181	0.147	0.222	0.099	0.105	1.057	84,855	0.683	0.565	0.825	0.097	0.096	0.988	93,581
ipw_boost	0.200	0.153	0.263	0.133	0.092	0.691	95,988	0.695	0.536	0.902	0.132	0.085	0.643	94,383
gps_mn	0.185	0.147	0.234	0.114	0.103	0.898	88,365	0.644	0.519	0.799	0.105	0.095	0.901	86,058
cl_mn	0.190	0.105	0.346	0.341	0.316	0.926	95,888	0.691	0.483	0.987	0.186	0.183	0.983	94,584
	0.191	0.127	0.286	0.210	0.210	1.003	95,687	0.695	0.506	0.954	0.162	0.162	1.001	95,186

Figure 4.13:
Case D

4.4 Discussion

In the scientific literature, there is a paucity of papers that concern the causal estimations of multiple treatments effects on a survival outcome in observational studies. It is in fact uncommon to find reports about the application of propensity score methods to compare more than two therapies.

This second part of the thesis has then been dedicated to the building of a simulation study to test the performances of propensity score models with three treatments. In fact, we aim for a validation of the methods previously utilised. Moreover, we believe that this topic is of general interest, since many clinical studies involve more than two therapies. It's possible to distinguish two main approaches: the first considers the treatments all together, while the second, that we called "classical", divides the analysis into two sub-problems, one concerning the comparison of the anatomic and the semi-anatomic resections and the other concerning the comparison of the anatomic and the wedge resections. The latter approach is in this way traced back to the binomial treatments setting.

In particular, we utilise the propensity score matching for the classical approach, as we did in the analysis of the large HCC dataset. It gives results that are overall good, except for the fact that they are characterized by very wide confidence intervals. Therefore, we have verified that even the classical approach is effective in multi-treatment settings and it has the advantage of utilising binary treatments methods that have already been solidly tested in disparate cases.

Also the IPW method hasn't been created ad hoc for the multi-exposure problems: it has actually been extended to include them. In fact, it still considers only one entry of the generalised propensity score vector, differently from the GPS-CDF model. Anyway, this fact doesn't impede the IPW multinomial method from working well. Quite the opposite, it is the method that performs better than all the others, like in the study case: it makes almost unbiased estimations of \mathbf{HR} with strict confidence intervals, it has satisfying percentages of coverage and its construction is pretty simple and not expensive. Its main flaw, shared with IPW gbm, is the fact that it considerably underestimates the sampling variability of the estimated $\log(\mathbf{HR})$.

The IPW generalised boosted model, instead, works even worse than the adjusted Cox model. As we've already pointed out, it is effective with non-linear relationships between covariates and treatments, but it is the opposite otherwise, like with these simulations. Moreover, the gbm regression is substantially more expensive than the multinomial logistic regression. Its application is then worthy with more complicated cases.

A similar conclusion can be stated for the GPS-CDF method. It doesn't make bad estimations of the \mathbf{HR} and has very good percentages of coverage, but its confidence intervals are way too large. Because of them, the uncertainty associated to \hat{HR}_s and \hat{HR}_w is too high. Hence, even GPS-CDF model should be more suitable to other cases, maybe with more than three treatments.

What should be finally highlighted is the fact that, in our results, changing the value of only one between HR_s and HR_w seems not to influence the estimations of the other hazard ratio. That makes us suppose that the performance of propensity score methods

for the estimation of HR_s is independent from the value of HR_w and vice versa. Nevertheless, a question arises: what if we considered more treatments? It is possible that a relationship between the performance of the methods and the value of the hazard ratios then emerges.

After all, in these simulations we have considered only three treatments. In these cases, we can conclude that both IPW model and classical approach still perform well, while GPS-CDF method, set up specifically for multi-treatments settings, isn't the best choice. Anyway, it would be interesting to extend this analysis to include more exposures. The paper [31] already shows that with the increasing of the number of treatments the IPW performances worsen. Then, further studies could examine better this aspect with new applications and also investigate how the classical approach would work. For sure, the more are the treatments considered, the more chaotic is dividing the problem into binary sub-problems. Moreover, it would be interesting verifying if the GPS-CDF performances enhance compared to the other methods.

Since the complexity of the analysis increases with the number of considered treatments, it is still important to remember that it is necessary to make a trade off between the completeness and the practical feasibility of the methods.

In conclusion, note that it is more common focusing on a low number of treatments - such as two or three -, than on too many of them.

5. Conclusion

This thesis is composed by two main parts:

1. The comparison of resection techniques used for large hepatocellular carcinoma. We utilised causal inference methods to define if one of the three following categories of surgical operations - anatomic, semi-anatomic and wedge resections - has better effects than the others with large HCC, for both overall and relapse-free survival.
2. The analysis of propensity score methods performances in three-treatments settings. We built a simulation study with the Monte Carlo method, generating 4 datasets analogous to the one previously studied and composed by 1000 instances each. Then, we applied five of the methods utilised before to these new cases, with the objective of testing their performances and verifying if they still work well with non-binary exposures.

In the analysis of the case of study, we had to face as main issue the limited size of the dataset. Inevitably, the privileged method is the one that suffers less from the data dimension, which is the inverse probability weighting using the propensity score. Unfortunately, we didn't find any other model as competitive as it.

Even the propensity score matching shows an acceptable performance, but it generates and uses very restricted matched datasets. In fact, we first had to consider two subsets of the original dataframe - a set composed by patients treated with anatomic or semi-anatomic resections and another composed by patients treated with anatomic or wedge resections - and, then, we utilised the matching process to form the matched sets, further reducing the final size by discarding the unmatched individuals.

The generalised propensity score cumulative distribution function model is even more affected by the small dimension of the dataset. In fact, it divides the original data into three disjoint matched sets - anatomic/semi-anatomic, anatomic/wedge and semi-anatomic/wedge -, greatly reducing the considered matched samples sizes.

Given this issue, since we took into consideration only the set of patients affected by large HCC, one interesting solution could be the expansion of the dataset of study and the development of new analyses.

In the end, we conclude that, with this data, there is no significant marginal treatment effects for both the overall and the relapse-free survival. In other words, there isn't any resection technique that appears to be better than the others for curing the large HCC. By contrast, the adjusted Cox model estimates for the overall survival a considerable difference between the effects of the anatomic and the wedge operations. In particular, the latter seems to perform better.

Also the ATT estimated by the propensity score matching method for the overall survival in the comparison of the anatomic and the wedge resections shows a substantial difference between them. The obtained result indicates that if patients operated with

the wedge method had been instead subjected to the anatomic intervention, the survival outcomes would have been worse.

Taking all these results in a whole and considering the fact that even from a clinical point of view the consequences of the surgery should depend also on the technique used, we can assume that a difference in the treatments effects exists, but not considering only the tumour dimension. Conversely, it can be detected on the basis of the overall patients' characteristics: it's likely that each individual has his optimal treatment, which could potentially be targeted on his health conditions.

In light of these observations, further studies could focus on the research of personalized therapies. In particular, it would be interesting to develop two models: the first with the intent of predicting the time of event giving in input the patients' general characteristics, while the second would be a classification model that associates each subject to the optimal resection technique for them. To be constructed, the latter needs the prediction model. Moreover, to reach good performances, it requires large quantity of data, in-depth analyses and much effort. Nevertheless, an efficient and accurate classifier for this scope could bring to great benefits.

Now, let's focus on the simulation study. It confirms the efficiency of the IPW multinomial method, which is still the model with the best performances, and of the propensity score matching method. Even this time the GPS-CDF model presents an important flaw, which is the fact that it estimates too wide confidence intervals. Hence, its results are associated to high uncertainties and are, then, little meaningful.

We conclude that, with three treatments, classical approaches or the IPW method extended to the multi-treatment setting are still valid and a better choice than GPS-CDF. The fact that they don't even use the generalised propensity score doesn't compromise their results. This conclusion is relevant, since the methods which are simpler and less computationally expensive are also the most effective.

The next challenge could be the study of propensity score methods considering more than three treatments. It is anyway important to remember that comparing too many therapies would considerably complicate the computation and their evaluation. Then, it is likely that, in the majority of cases, researchers would take into account a relatively low number of treatments.

In the end, in the first part, we rejected the hypothesis that one of the considered resection techniques works better than the others with large HCC. Nevertheless, our solution made new challenging research tasks emerge.

Instead, from the simulation study, we can infer that "classical" propensity score methods are still unbiased in the comparison of three treatments. Therefore, the use of more complicated methods built specifically for multi-treatment settings should be limited to cases characterized by numerous therapies. However, it is uncommon that studies involve many treatments: indeed, adopting a simplification to reduce their number during the problem modelling is preferred and convenient.

In conclusion, further studies could focus on models performances with more than three exposures, observing how they change with the increasing of the number of treatments.

Acknowledges

This thesis has been possible thanks to the support and the patience of the Dr. Davide Paolo Bernasconi, who guided me during the study and the writing. A special thank goes also to the Dr. Giulia Capitoli, for her kindness and her helpfulness.

This work is just the last step of five intense years spent in the University of Bicocca. It has been a gratifying path, which seems to have passed in a flash. At the end of the bachelor's degree, I had the impression that I often escaped from the academic life. Nevertheless, I have the luck of having met wonderful people, with whom, during the Master, I did the little things that, before, I had regretted not to have done. Like finally visiting the Hangar Bicocca, thanks to Matteo.

I would like to express my gratitude even to Klagenfurt, a place whose existence was unknown to me just two years ago, before the Erasmus study+ program selected me to go there.

There are many things that I didn't learn during my semester there - like cooking, cleaning, tidying my room or using the washing machine properly -, but the most important thing is what I shared with whom I met there. Even if sometimes it was only a room too small, or a dish of pasta (*literally* the same dish).

In particular, I considered myself very lucky to share my graduation day with Camilla and Iman, who spent the last semester with me in Klagenfurt.

Last but not least, my gratitude goes also to my parents and to my family for giving me their trust and the opportunity of accomplishing this important goal.

To my portable computer that loyally safeguarded my codes, even after my mistreatments.

To my old friends, who shared the years of school with me and strangely haven't decided to abandon me, yet.

And to Ruben, who always had to support my worries and my mood swings and who, after all, stayed beside me, no matter if physically near or far.

References

- [1] A. B. Hill, “The environment and disease: association or causation?” 1965.
- [2] K. M. Fedak, A. Bernal, Z. A. Capshaw, and S. Gross, “Applying the bradford hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology,” *Emerging themes in epidemiology*, vol. 12, no. 1, pp. 1–9, 2015.
- [3] K. J. Rothman and S. Greenland, “Causation and causal inference in epidemiology,” *American journal of public health*, vol. 95, no. S1, pp. S144–S150, 2005.
- [4] K. Suresh, “An overview of randomization techniques: an unbiased assessment of outcome in clinical research,” *Journal of human reproductive sciences*, vol. 4, no. 1, p. 8, 2011.
- [5] D. A. Grimes and K. F. Schulz, “An overview of clinical research: the lay of the land,” *The Lancet*, vol. 359, no. 9300, pp. 57–61, 2002. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0140673602072835>
- [6] C. Kartsonaki, “Survival analysis,” *Diagnostic Histopathology*, vol. 22, no. 7, pp. 263–270, 2016, mini-Symposium: Medical Statistics. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1756231716300639>
- [7] M. Sestelo, *A short course on Survival Analysis applied to the Financial Industry*. bookdown.org, 2017. [Online]. Available: https://github.com/sestelo/sa_financial
- [8] D. G. Kleinbaum, M. Klein *et al.*, *Survival analysis: a self-learning text*. Springer, 2012, vol. 3.
- [9] P. C. Austin, D. S. Lee, and J. P. Fine, “Introduction to the analysis of survival data in the presence of competing risks,” *Circulation*, vol. 133, no. 6, pp. 601–609, 2016.
- [10] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
- [11] R. Charan, “The cox proportional hazards model,” accessed: 09/09/2022. [Online]. Available: <https://towardsdatascience.com/the-cox-proportional-hazards-model-35e60e554d8f>
- [12] “Logrank test,” accessed: 09/09/2022. [Online]. Available: <https://web.stanford.edu/~lutian/coursepdf/unitweek3.pdf>
- [13] J. M. Bland and D. G. Altman, “The logrank test,” *Bmj*, vol. 328, no. 7447, p. 1073, 2004.

- [14] A. Rothman, “Mathematical statistics — rigorous derivations and analysis of the wald test, score test, and likelihood ratio test,” accessed: 09/10/2022. [Online]. Available: <https://towardsdatascience.com/mathematical-statistics-a-rigorous-derivation-and-analysis-of-the-wald-test-score-test-and-6262bed53c55>
- [15] P. W. Holland, “Statistics and causal inference,” *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.
- [16] D. B. Rubin, “Causal inference using potential outcomes,” *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005. [Online]. Available: <https://doi.org/10.1198/016214504000001880>
- [17] M. M. Garrido, B. Dowd, P. L. Hebert, and M. L. Maciejewski, “Understanding treatment effect terminology in pain and symptom management research,” *Journal of pain and symptom management*, vol. 52, no. 3, pp. 446–452, 2016.
- [18] P. C. Austin, “The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments,” *Statistics in medicine*, vol. 33, no. 7, pp. 1242–1258, 2014.
- [19] —, “An introduction to propensity score methods for reducing the effects of confounding in observational studies,” *Multivariate Behavioral Research*, vol. 46, no. 3, pp. 399–424, 2011, pMID: 21818162. [Online]. Available: <https://doi.org/10.1080/00273171.2011.568786>
- [20] A. Remiro-Azócar, A. Heath, and G. Baio, “Conflating marginal and conditional treatment effects: Comments on “assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study”,” *Statistics in Medicine*, vol. 40, no. 11, pp. 2753–2758, 2021.
- [21] T. Grigg, “Simpson’s paradox and interpreting data,” 2018, accessed: 09/09/2022. [Online]. Available: <https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765>
- [22] P. R. Rosenbaum and D. B. Rubin, “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
- [23] M. S. Ali, D. Prieto-Alhambra, L. C. Lopes, D. Ramos, N. Bispo, M. Y. Ichihara, J. M. Pescarini, E. Williamson, R. L. Fiaccone, M. L. Barreto *et al.*, “Propensity score methods in health technology assessment: principles, extended applications, and recent advances,” *Frontiers in pharmacology*, p. 973, 2019.
- [24] D. W. Brown, S. M. DeSantis, T. J. Greene, V. Maroufy, A. Yaseen, H. Wu, G. Williams, and M. D. Swartz, “A novel approach for propensity score matching and stratification for multiple treatments: Application to an electronic health record-derived study,” *Statistics in medicine*, vol. 39, no. 17, pp. 2308–2323, 2020.

- [25] X. S. Gu and P. R. Rosenbaum, “Comparison of multivariate matching methods: Structures, distances, and algorithms,” *Journal of Computational and Graphical Statistics*, vol. 2, no. 4, pp. 405–420, 1993.
- [26] P. C. Austin, “A comparison of 12 algorithms for matching on the propensity score,” *Statistics in medicine*, vol. 33, no. 6, pp. 1057–1069, 2014.
- [27] N. Greifer, “ *MatchIt* r package,” May 2022. [Online]. Available: <https://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf>
- [28] S. Xu, C. Ross, M. A. Raebel, S. Shetterly, C. Blanchette, and D. Smith, “Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals,” *Value in Health*, vol. 13, no. 2, pp. 273–277, 2010.
- [29] N. Greifer, “ *WeightIt* r package,” June 2022. [Online]. Available: <https://cran.r-project.org/web/packages/WeightIt/WeightIt.pdf>
- [30] K. Imai and D. A. Van Dyk, “Causal inference with general treatment regimes: Generalizing the propensity score,” *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 854–866, 2004.
- [31] S. Yang, G. W. Imbens, Z. Cui, D. E. Faries, and Z. Kadziola, “Propensity score matching and subclassification in observational studies with multi-level treatments,” *Biometrics*, vol. 72, no. 4, pp. 1055–1065, 2016.
- [32] P. E. college of Science, “7.2 - probability mass functions,” accessed: 09/09/2022. [Online]. Available: <https://online.stat.psu.edu/stat414/lesson/7/7.2>
- [33] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.
- [34] D. W. Brown, T. J. Greene, and S. M. DeSantis, “ *GPSCDF* r package,” 03 2019. [Online]. Available: <https://cran.r-project.org/web/packages/GPSCDF/index.html>
- [35] B. Greenwell, B. Boehmke, J. Cunningham, and G. Developers, “ *gbm* r package,” 08 2022. [Online]. Available: <https://cran.r-project.org/web/packages/gbm/gbm.pdf>
- [36] “Probability vs odds,” accessed: 09/09/2022. [Online]. Available: <https://towardsdatascience.com/probability-vs-odds-f47fbc6789f4>
- [37] ibm, “What is logistic regression?” accessed: 09/09/2022. [Online]. Available: <https://www.ibm.com/topics/logistic-regression>
- [38] A. Schüppert, “Binomial (or binary) logistic regression,” in *Statistics Seminar, Spring*. Citeseer, 2009.
- [39] E. García-Portugués, *Notes for Predictive Modeling*. bookdown.org, 2017, last update: 25/04/2022. [Online]. Available: <https://bookdown.org/egarpor/PM-UC3M/>

- [40] B. Ripley and W. Venables, “*nnet* r package,” 01 2022. [Online]. Available: <https://cran.r-project.org/web/packages/nnet/nnet.pdf>
- [41] bccvl, “Generalized boosting model,” accessed: 09/09/2022. [Online]. Available: <https://support.bccvl.org.au/support/solutions/articles/6000083212-generalized-boosting-model#:~:text=These%20models%20are%20a%20combination,selected%20using%20the%20boosting%20method.>
- [42] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.
- [43] ibm, “Boosting,” accessed: 09/09/2022. [Online]. Available: <https://www.ibm.com/cloud/learn/boosting#:~:text=Boosting%20is%20an%20ensemble%20learning,the%20weaknesses%20of%20its%20predecessor.>
- [44] N. Greifer, “Propensity score weighting using generalized boosted models,” accessed: 09/09/2022. [Online]. Available: https://ngreifer.github.io/WeightIt/reference/method_gbm.html
- [45] W. H. Organization, “Cancer,” February 2022. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/cancer>
- [46] J. M. Llovet, R. K. Kelley, A. Villanueva, A. G. Singal, E. Pikarsky, S. Roayaie, R. Lencioni, K. Koike, J. Zucman-Rossi, and R. S. Finn, “Hepatocellular carcinoma.” *Nat Rev Dis Primers*, pp. 6–6, 2021.
- [47] cancer.org, “Liver cancer risk factors,” accessed: 09/09/2022. [Online]. Available: <https://www.cancer.org/cancer/liver-cancer/causes-risks-prevention/risk-factors.html#:~:text=Worldwide%2C%20the%20most%20common%20risk,many%20parts%20of%20the%20world.>
- [48] S.-C. Chuang, C. La Vecchia, and P. Boffetta, “Liver cancer: descriptive epidemiology and risk factors other than hbv and hcv infection,” *Cancer letters*, vol. 286, no. 1, pp. 9–14, 2009.
- [49] Hercoles, “Hepatocarcinoma recurrence on the liver study – Fase1,” 25 November 2018. [Online]. Available: https://www.hercolesgroup.eu/_files/ugd/edc0e3_34b5cd35a65d4d0a83db66998a76c7c1.pdf
- [50] S. Famularo, M. Donadon, F. Cipriani, F. Ardito, F. Carissimi, P. Perri, M. Iaria, T. Dominioni, M. Zanella, S. Conci *et al.*, “Hepatocellular carcinoma surgical and oncological trends in a national multicentric population: the hercoles experience,” *Updates in surgery*, vol. 72, no. 2, pp. 399–411, 2020.
- [51] K. Labadie, K. M. Sullivan, and J. O. Park, “Surgical resection in hcc,” in *Liver Cancer*. IntechOpen, 2018.

- [52] V. W. Keng, D. A. Largaespada, and A. Villanueva, “Why men are at higher risk for hepatocellular carcinoma?” *Journal of hepatology*, vol. 57, no. 2, pp. 453–454, 2012.
- [53] D. J. Doyle, A. Goyal, P. Bansal, and E. H. Garmon, “American society of anesthesiologists classification,” in *Statpearls [internet]*. StatPearls Publishing, 2021.
- [54] M. E. Charlson, D. Carrozzino, J. Guidi, and C. Patierno, “Charlson comorbidity index: a critical review of clinimetric properties,” *Psychotherapy and Psychosomatics*, vol. 91, no. 1, pp. 8–35, 2022.
- [55] K. Shakerdge, “What are meld and child-pugh scores?” accessed: 09/09/2022. [Online]. Available: <https://www.webmd.com/hepatitis/meld-score-for-liver-disease>
- [56] M. S. Williams, “What are platelets and why are they important?” accessed: 09/09/2022. [Online]. Available: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/what-are-platelets-and-why-are-they-important>
- [57] L. Zhou, J.-A. Rui, W.-X. Zhou, S.-B. Wang, S.-G. Chen, and Q. Qu, “Edmondson-steiner grade: A crucial predictor of recurrence and survival in hepatocellular carcinoma without microvascular invasion,” *Pathology-Research and Practice*, vol. 213, no. 7, pp. 824–830, 2017.
- [58] E. Ünal, İ. S. İdilman, D. Akata, M. N. Özmen, and M. Karçaaltıncaba, “Microvascular invasion in hepatocellular carcinoma,” *Diagnostic and Interventional Radiology*, vol. 22, no. 2, p. 125, 2016.
- [59] E.-S. Cho and J.-Y. Choi, “Mri features of hepatocellular carcinoma related to biologic behavior,” *Korean journal of radiology*, vol. 16, no. 3, pp. 449–464, 2015.
- [60] P. Hermanek and C. Wittekind, “The pathologist and the residual tumor (r) classification,” *Pathology-Research and Practice*, vol. 190, no. 2, pp. 115–123, 1994.
- [61] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, “Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls,” *Bmj*, vol. 338, 2009.
- [62] S. van Buuren, “*mice* r package,” November 2021. [Online]. Available: <https://cran.r-project.org/web/packages/mice/mice.pdf>
- [63] T. J. Greene, S. M. DeSantis, D. W. Brown, A. V. Wilkinson, and M. D. Swartz, “A machine learning compatible method for ordinal propensity score stratification and matching,” *Statistics in medicine*, vol. 40, no. 6, pp. 1383–1399, 2021.
- [64] N. Metropolis, “The beginning of the monte carlo method,” in *Los Alamos Science*, 1987.

- [65] N. Metropolis and S. Ulam, “The monte carlo method,” *Journal of the American statistical association*, vol. 44, no. 247, pp. 335–341, 1949.
- [66] RDocumentation, “rnorm: Normal distributions on special spaces,” accessed: 09/09/2022. [Online]. Available: <https://www.rdocumentation.org/packages/compositions/versions/2.0-4/topics/rnorm>
- [67] —, “Binomial: The binomial distribution,” accessed: 09/09/2022. [Online]. Available: <https://www.rdocumentation.org/packages/stats/versions/3.3/topics/Binomial>
- [68] —, “invlogit: Logistic and inverse logistic functions.” [Online]. Available: <https://www.rdocumentation.org/packages/arm/versions/1.12-2/topics/invlogit>
- [69] A. Nava, “Thesis: HCC study,” September 2022. [Online]. Available: <https://github.com/anna-38/Thesis>
- [70] P. C. Austin and E. A. Stuart, “Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies,” *Statistics in Medicine*, vol. 34, no. 28, pp. 3661–3679, 2015. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6607>
- [71] J. M. Llovet, R. K. Kelley, A. Villanueva, A. G. Singal, E. Pikarsky, S. Roayaie, R. Lencioni, K. Koike, J. Zucman-Rossi, and R. S. Finn, “Hepatocellular carcinoma,” *Nature reviews. Disease primers*, vol. 34, no. 28, pp. 3661–3679, 2021. [Online]. Available: <https://www.nature.com/articles/s41572-020-00240-3>
- [72] D. Machin and M. J. Campbell, *The design of studies for medical research*. John Wiley & Sons, 2005.
- [73] S. V. Deo, V. Deo, and V. Sundaram, “Survival analysis—part 2: Cox proportional hazards model,” *Indian journal of thoracic and cardiovascular surgery*, vol. 37, no. 2, pp. 229–233, 2021.
- [74] M. M. Garrido, J. Lum, and S. D. Pizer, “Vector-based kernel weighting: A simple estimator for improving precision and bias of average treatment effects in multiple treatment settings,” *Statistics in medicine*, vol. 40, no. 5, pp. 1204–1223, 2021.
- [75] P. C. Austin and T. Schuster, “The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study,” *Statistical methods in medical research*, vol. 25, no. 5, pp. 2214–2237, 2016.
- [76] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.