ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus School of Economics

Master Thesis Econometrics & Management Science

Impact of waiting time on post-transplant survival for patients with Hepatocellular Carcinoma

An instrumental variable analysis

Berend Robert Beumer

381166

Supervisor:

Dr. AA Naghi

External supervisor:

Prof. Dr. JNM IJzermans

Second assessor:

Dr. E O'Neill

January 20, 2022

The views stated in this thesis are those of the author and not necessarily those of Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

A donor liver for transplantation (LT) is not always immediately available for eligible patients with Hepatocellular Carcinoma (HCC). Therefore, patients are placed on a waiting list. In this thesis we will estimate the causal effect of waiting time on post-operative overall survival for transplanted HCC patients. The cost of waiting is important for policy decisions regarding for example, the test-of-time, the waiting list upper bound, and the choice between multiple treatments. Establishing the causal effect of waiting time is, however, complicated due to multiple confounding and selection biases. Therefore, we have provided a comprehensive discussion on the causal inference techniques supporting our results. In this thesis we present the first causal graph in the field, which represents our model of reality and simultaneously explicates our assumptions. We will show that blood group can be used as a valid instrumental variable. Application of instrumental variable analysis in the context of a continuous treatment and time-to-event outcome is, however, complicated due to the non-collapsibility of the proportional hazard model. We will show how the additive hazards model can be used as the second stage in the two-stage predictor substitution framework. The analysis resulted in that waiting is harming post-transplant survival. More specifically, if a patient waits 12 instead of 2 months, we estimated a corresponding loss of 3.37 years in terms of the median post-transplant survival. Using interaction terms, we did not find significant heterogeneity of the treatment effect. Lastly, we performed a cost-analysis for the test-of-time, and we provided the conditions under which implementation of this test would result in a survival benefit.

Contents

1 Introduction

A donor liver for transplantation (LT) is not always immediately available for eligible patients with Hepatocellular Carcinoma (HCC), therefore patients are placed on a waiting list. How this waiting list should be organized continues to be a topic of fierce debate. In the past two decades high quality research was aimed at answering major questions like: which subset of patients should be placed on the waiting list? What is the best way to treat patients prior to transplantation? And, how do we prioritize high risk patients to reduce waiting list mortality? Continued research on these major topics will keep advancing the discussion on how to best organize clinical care. Although there is a growing consensus about these major topics, a small question fundamental to organizing the waiting list was never correctly answered. Namely, how much does waiting longer cost a patient with HCC in terms of post-transplant survival?

Knowing the degree of harm caused by time itself is essential to determine an upper bound, or to confidently expand the waiting list. This is especially important for patients with HCC, as these patients become increasingly more ill on the waiting list. Additionally, the use of the so-called test-of-time is increasingly promoted. The rationale of this policy is that in the perioperative observation period patients with the most aggressive forms of cancer are filtered out, in order to better allocate the scarce donor livers and increase the average post-transplant survival. However, the merit of this allocation policy depends on whether the harm of the additional waiting time for the full population can be offset with the better allocation of a fraction of the livers. Lastly, knowing the harm of waiting can aid transplant clinicians to recommend treatment for patients with multiple options. Patients with more treatment options are often in an earlier stage with good liver function. This also leads to a low ranking in the waiting list. Knowing the harm of waiting enables clinicians to identify a threshold at which resection is the better strategy compared to bridging therapy and eventually transplantation.

Establishing the causal effect of waiting time is, however, complicated by the prioritization schemes. To reduce the waiting list mortality, the waiting list is currently organized such that the patients with the lowest liver function are transplanted first. Therefore, a naive survival comparison of the patients transplanted within 6 months versus those transplanted after more than 6 months, or a variation thereof, is guaranteed to result in downward biased estimates. It would seem that waiting is beneficial. A further complicating factor is that socioeconomic mechanisms play an important role. It is known that patients with lower socioeconomic status are more often inactive on the waiting list due to, for example, problems with their insurance, incomplete screening, or medical non-compliance. Additionally, lower frequency of check-up visits leads to a slower escalation of their priority in the waiting list if their liver function decreases. Even though a randomized controlled trial is the golden standard to avoid these biases, in this case it is less preferred due to ethical considerations and the time and expenditures involved. Quasi experimental studies are therefore the best alternative.

In this case, the random assignment of blood groups creates a natural experiment, with treatment arms that are largely equal in composition of (unobserved) confounders but differ with regard to waiting time. Key is that the assignment of a blood group is determined at conception and follows the fundamental laws of inheritance introduced by Mendel (Haycock et al., 2016). In other words, patients or doctors cannot choose the blood group. Additionally, the blood group itself is unlikely to directly influence the severity of the HCC, liver function, complications, or survival (Oral and Sahin, 2019). The blood groups, however, do directly influence a patient's time to transplant. This is caused by the fact that patients with blood group AB can accept a donor organ from blood group A, B, AB, or O. While patients with blood group O can only receive a donor organ from blood group O. Due to this asymetry, the blood group could be used as an instrument for waiting time in an instrumental variable analysis. Hence, the research aim of this thesis is to exploit the natural experiment created by blood groups and estimate the causal effect of waiting time on postoperative overall survival (OS) for transplanted HCC patients.

2 Literature review

Over the past two decades several research groups, among which Freeman Jr and Edwards (2000), Schlansky et al. (2014), Salvalaggio et al. (2015), have investigated the effect of waiting time on post-transplant survival. They concluded that waiting longer was beneficial or harmless for transplant candidates with HCC. However, they did not use a causal model to show their assumptions, which led to a wide variety of conditioning sets. It remains unclear if they sufficiently accounted for selection or confounding bias. Ultimately, their consensus led in 2015 to the implementation of the mandatory 6-month waiting policy for HCC patients. In the new policy, HCC patients were placed on the waiting list ranked only based on their MELD-Na score, describing their liver function, and were only assigned exception points for HCC after 6 months of waiting (HRSA/OPTN, 2018).

In addition to these associational studies, several studies used causal inference to investigate the effect of the waiting time. Recently, Nagai et al. (2020) used a before-after design to investigate the introduction of the mandatory 6-month waiting period for HCC patients. Their primary aim was to investigate waiting list mortality and dropout, however, they also briefly investigated post-transplant outcomes. The authors reported that the policy change had no effect on post-transplant mortality. However, a few limitations, inherent to the before-after study design, made interpretation of their results difficult. The before-after study design was likely biased by trends in the post-transplant survival. As the inclusion period spanned more than 5 years, and improvements in terms of imaging and clinical care were not controlled for, the estimates are likely biased toward the null. Further, apart from changes in the treatment technique, the composition of the before and after group changed. The authors anticipated a higher dropout rate and better post-transplant outcomes, due to the selection mechanisms similar to those used in the test-of-time. However, their results showed both lower dropout and lower mortality rates in the group which experienced the mandatory 6-month waiting. Nagai et al. stated that after the policy change doctors became more reluctant to place patients with advanced disease on the waiting list, knowing that exception points are only assigned after 6 months of waiting. In their analysis they did not address the multitude of these selection processes. Furthermore, they acknowledged that a longer follow-up was needed to study post-transplant survival in more detail.

An alternative causal inference technique is to exploit the exogenous variation in a treatment created by an instrumental variable. The research by Everhart et al. (1997) was the first to analyse the survival of patients stratified by blood group. The analysis was, however, limited to stratification. Several other aspects of the study were problematic for identifying the cost of waiting in terms of survival. Most importantly, the study described patients treated two decades ago (1990-1993) and the indications for transplantation and the waiting list policy have changed since then. For this reason their study also did not include patients with HCC, and is therefore not representative for our population. Further, their sample size of 673 patients was limited. And, the logistic model they used to study the time-to-event data was likely biased due to the more frequent censoring of patients with a longer survival. Lastly, the interpretation of their results was further hindered, by their choice of endpoint. They modeled death within 2 years after listing, spanning a period of both pre- and post-transplantation, making cross reference with other literature hard. Later, Howard (2000) expanded on the work of Everhart and used blood groups as an instrument in a two-stage regression approach. However, this analysis did not focus on post-transplant survival, but on graft failure within 3 months post surgery. Instead of addressing the time-to-event data structure they heuristically employed a logistic model in the second stage.

In this thesis we will extend this empirical research and build on the instrumental variable analysis (IV) which was first introduced by Wright (1928) and later popularized by publications of Angrist (1990), Angrist and Keueger (1991) and Angrist and Lavy (1999), among others. The applications of these IV methodologies, however, primarily focus on binary instruments, binary treatments and continuous outcome variables to identify the treatment effect for compliers Angrist et al. (1996). Earlier work by Theil (1953) and Basmann (1957) generalized the Wald estimator (Wald, 1940) to the two-stage regression framework allowing for categorical instrumental variables and continuous treatments. With respect to the outcome, for survival analysis the proportional hazards model is by far the most popular survival analysis technique. However, all resulting coefficients are affected by omitted variable bias. Attempts by MacKenzie et al. (2014) to justify the use of the cox model in an IV setting, only identified the treatment effect under the implausible condition that confounding acts additively on the hazard. Tchetgen et al. (2015) suggested that the application of the cox model as a second stage can be heuristically applied in the limited context of rare events. In that case the non-collapsibility is negligible. Earlier, Robins and Tsiatis (1991) have developed a IV approach using G-estimation and an accelerated failure time model. Although, solutions to the estimating equations do not always exist. Furthermore, the method lacks efficiency. To retain unconfoundedness, patients of which the event time was observed are artificially censored. As a collapsible alternative, Tchetgen et al. (2015) proposes the additive hazards model, which was first introduced by Aalen (1980). Asymptotic properties were proven and verified in simulation studies by Li et al. (2015). Therefore, the additive hazards model will be used in this thesis.

In addition to the average treatment effect (ATE) it is of interest to study the heterogeneity of the treatment effect. Over the past decade several methods have been suggested, most focus on partitioning the covariate space and estimating the average treatment effect within each of the subsets. The challenge is, however, how to best split the space, how to attain smooth transitions between adjacent subsets, and how to perform inference with regard to whether the subset treatment effect is different from zero or whether the subset treatment effect is different from the ATE. Traditionally quantiles were used, however, more recently honest causal trees by Athey and Imbens (2016) are proposed as a more flexible alternative. Causal random forests are used to further extended the causal trees and obtain a smooth estimate Wager and Athey (2018). Tree methods are, however, developed and most commonly used for binary treatments and continuous outcomes. Generalizations to continuous treatments have been proposed by Athey et al. (2019) which extend the causal forest framework to all generalised method of moments estimators (GMM). Recently causal survival forrest have been developed by Cui et al. (2020). However, to the best of our knowledge, no research group has yet combined instrumental variable analysis with causal survival forrests. Other often proposed smoothing methods are k-nearest neighbors, kernel smoothing, or local linear regression. These are, however, exposed to the curse of dimensionality, and inference is poorly developed making it unsuitable for this thesis. Yet, interaction terms can be applied in the instrumental variable and time-to-event outcome context. At the cost of assuming linear changes in the treatment effect, interaction terms do provide a natural way to compare the effect in subgroups to the ATE.

3 Data

In this section we describe the origin of the data that is used for this research, what in- and exclusion criteria were followed, how the files constituting the data set are structured, and how the data was preprocessed.

The data for this research originated from the United Network for Organ Sharing (UNOS) which is the scientific organization of the Organ Procurement and Transplantation Network (OPTN) that coordinates all transplantations in the United States of America. After appropriate ethics approval was obtained from the medical ethics committee at the Erasmus Medical Center (MEC-2020-0779) they provided the data. The dataset contained the records of 318,004 patients receiving liver transplantation in the period spanning from December 1985 to December 2020. For this research only a subset of the data was used. If patients received more than one liver transplantation, only the first transplantation was included in the analysis. Furthermore, only listings in the period from 2000-2019 were selected to exclude experimental treatments in the period before 2000 and to allow for adequate follow-up for those listed in 2019. The UNOS data set is structured into five separate files describing the transplantation process in varying levels of detail. The file *liver data* is the main file, which is organized per listing and contains information regarding the transplantation. The other files are supplementary to the main file. More specifically, *liver* *wlhistory* contains information on visits during the waiting period, the *liver exception* file contains information on tumor characteristics obtained from radiology prior to transplantation, the *liver explant* file provides information obtained from histopathological examination after the transplantation, and lastly the *liver followup* file contains information indexed per follow-up visit after the transplantation on graft failure and post-transplant survival. For this research, the file *liver data* was the main source of information, and was extended with relevant data from the other files. All together the files contain 652 variables, of which only those relevant to answer our research question were selected and are presented in Table 1. Each row in the file *liver data* represents a listing record. Some patients were transferred multiple times between hospitals and were, at each hospital, re-enlisted. To account for these multiple listings, that are always only partially complete, appropriate aggregations were made for each of the variables. Most importantly, the waiting time was re-calculated as the time difference between the date of first listing, out of all listings with indication HCC, and the date of the first transplantation or last follow-up. In addition, the distributions of all variables were inspected, and several outliers were removed. An exhaustive list of the modifications can be found in Appendix A.

Table 1: Relevant variables

Variable name	**Variable description**	**Variable name**	**Variable description**
G	Blood group	$Tum\ num^{list}$	Tumor number at listing
W	Waiting time in days	$Tum\ size^{list}$	Tumor size at listing
T	Time to death or censoring	AFP^{list}	*Log*_10(AFP) at listing
$Gender$	Gender	$Enceph$	Encephalopathy
Age	Age at listing	$Ascites$	Ascites
BMI	BMI at listing	$Cirrhosis$	Cirrhosis
$Functional\ status$	Functional status at listing	$ALBI^{lt}$	ALBI score at LT
$Life\ support$	Life support at listing	$MELD^{lt}$	MELD score at LT
$Educ$	Highest level of education	$Tum\ num^{lt}$	Tumor number last prior to LT
$Ethn$	Ethnicity category	$Tum\ size^{lt}$	Tumor size last prior to LT
$Payment\ type$	Primary projected payment type	$Tum\ size\ tot^{lt}$	Cumulative tumor size last prior to LT
$Region$	Region	AFP^{lt}	*Log*_10(AFP) last prior to LT
$MELD^{list}$	MELD score at listing	$dropout$	Waiting list dropout
$ALBI^{list}$	ALBI score at listing		

Missing data still remained after supplementing the variables in the main file with the information from the more detailed supplementary files. These missing values were addressed using multiple imputation. In this iterative procedure of modelling and predicting the missing values, predictive mean matching, linear, logistic, or multinomial logistic regression were used. To account for the sampling uncertainty related to the imputation, multiple versions of the imputed data set were made. In this research the fraction of missing data was 10%. Hence, following the advice of Graham et al. (2007), the missing values were imputed 20 times (m=20) with each imputation receiving 20 iterations of predicting the missing value and re-estimating the prediction model. After imputation, we evaluated that this number of iterations was sufficient to reach convergence as is shown in Appendix B. After repeating the primary analysis on each of the imputed data sets, the resulting estimates were pooled using the Rubin Rules (Rubin, 2004). The pooled point estimate was the mean:

$$\hat{\theta} = \frac{1}{m}\sum_{j=1}^{m}\theta_j.$$

In which m is the number of imputed data sets and θ_j is the parameter estimate in imputed data set j. The standard errors are based on the within and between sample variance of parameters in the multiple imputed data sets. The within sample variance V_W is calculated as the mean squared standard error in each of the imputed data sets:

$$V_W = \frac{1}{m}\sum_{j=1}^{m} SE_j^2.$$

The between imputation variance V_B is calculated as the average squared distance between the parameter estimate of a particular imputation, and the mean parameter estimate, over all imputations:

$$V_B = \frac{\sum_{j=1}^{m}(\theta_i - \bar{\theta})}{m-1}.$$

With these the SE_{pooled} can be calculated by taking the square root of V_T, which is calculated as follows:

$$V_T = V_W + V_B + \frac{V_B}{m}.$$

Further, significance testing of the null hypothesis θ_0 is performed using the regular Wald test statistic,

$$Wald_{pooled} = \frac{(\hat{\theta} - \theta_0)^2}{V_T}.$$

The test statistic follows a t-distribution with degrees of freedom calculated following the book of Van Buuren (2018). In this research we use the $df_{adjusted}$,

$$df_{adjusted} = \frac{df_{old} * df_{observed}}{df_{old} + df_{observed}}.$$

In which df_{old} and $df_{observed}$ are defined as:

$$df_{old} = (m-1)/(\frac{V_B + \frac{V_B}{m}}{V_T})^2$$

$$df_{observed} = \frac{(n-k)+1}{(n-k)+3} * (n-k)(1 - (\frac{V_B + \frac{V_B}{m}}{V_T})).$$

Where, n is the sample size and k the number of parameters.

4 Methodology

In the first subsection of the methodology we will explicitly describe, with the causal query, what we try to estimate in this thesis. Hereafter, we briefly describe the preliminaries of survival analysis which is used to introduce the additive hazards model. The collapsibility of the additive hazards model in relation to causal inference is inspected in more detail. Next, we will discuss the preliminaries of causal graphs in support of the causal graph we use for this research. Subsequently, we describe how we will account for the selection bias by means of inverse probability weighting. Where after, we consider instrumental variable analysis to address confounding bias. After which we will return to the causal query and establish the causal estimand. We conclude the methodology section with two post-hoc analysis. The first focuses how heterogeneity of the treatment effect can be investigated by means of sub setting, and interaction terms. The second focuses on a cost-analysis and identifies the conditions under which the test-of-time results in a net benefit.

4.1 Causal query

This thesis aims to estimate the ATE of waiting time on post-transplant survival. More formally defined, let W_i be the random variable, with realisation w_i, describing the waiting time between listing and transplantation for individuals i = 1, ..., n. Further, let $\tilde{T_i}$ denote the latent survival time after transplantation and $\tilde{C_i}$ the latent censoring time after transplantation. Ultimately, we observe $T_i = min(\tilde{T_i}, \tilde{C_i})$ and the survival status-indicator $\Delta_i = I(\tilde{T_i} \leq \tilde{C_i})$. In which we assume that $\tilde{T_i}$ and $\tilde{C_i}$ are conditionally independent given a (p+1)-dimensional covariate vector $L_i^T = [1, L_{i1}, L_{i2}, \ldots, L_{ip}]$. Besides, let $\lambda_i(t|L_i)$ denote the conditional hazard rate, such that $\lambda_i(t|L_i)dt$ describes the conditional probability of the event happening in the small time interval between t and $t + dt$ given that the event has not yet happened. Lastly, let W, T, Δ, L, $\lambda(t|L)$ be the n-dimensional variants of the above. Then the causal query can be written as the following hazard difference:

$$\lambda[t|do(W = w + 1)] - \lambda[t|do(W = w)]. \tag{1}$$

With regard to the contrast in expression 1, we will quantify the effect in this general format, rather than evaluate any specific waiting time contrast, at a particular time post-transplantation. Yet, to aid interpretation of the regression coefficients we will also evaluate the difference in 5-year post-transplant survival between those assigned waiting time of 2 months versus those assigned waiting time of 12 months:

$$P[T > 5\ year|\ do(W = 12\ months)] - P[T > 5\ year|\ do(W = 2\ months)].$$

In this contrast, the 5-year post-transplant survival is chosen as this is a common reference point in the medical literature. The waiting times of 2 and 12 months correspond approximately to the first quartile (55 days) and third quartile (361 days) of the waiting time distribution in the UNOS data set. Lastly, to inspect the merit of the test-of-time we will evaluate the cost of waiting as:

$$c = E[T|do(W = w)] - E[T|do(W = w + 1)].$$

Furthermore, we would like to point out that in the specification of the ATE we adopt the $do()$ notation to highlight that the treatment variable is set to a particular value and needs to be interpreted as if it was assigned in a randomized controlled trial. The variable W without the $do()$ operator implies that the waiting time is simply observed as it occurs, without external intervention. This distinction in notation between experimentation and observation gives the query its causal status.

4.2 Additive hazards model

In addition to the definition of the random variables in the previous section, let the distribution of T be described by probability density function $f(t)$ and cumulative probability distribution $F(t)$. Furthermore, the survival function describing the proportion of patients still alive at time t is defined as:

$$S(t) := 1 - F(t). \tag{2}$$

Furthermore, the hazard function that describes the rate at which deaths occur at time t and is defined as:

$$\begin{aligned} \lambda(t) :=& \lim_{h\to 0} \frac{P[t \leq T < t+h|T \geq t]}{h} \\ =& \frac{f(t)}{S(t)}. \end{aligned} \tag{3}$$

Integrating the hazard function with respect to time yields the cumulative hazard function: $\Lambda(t) := \int_0^t \lambda(u)du$. Lastly, as shown below the survival function and the cumulative hazard function have the following relationship: $S(t) = exp(-\Lambda(t))$.

$$\begin{aligned} S(t) &= 1 - F(t) \\ \frac{d}{dt}S(t) &= \frac{d}{dt}(1 - F(t)) \\ \frac{d}{dt}S(t) &= f(t) \\ f(t) &= -\frac{d}{dt}S(t) \quad \square \end{aligned} \tag{4}$$

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{s(t)} \\ &= -\frac{1}{s(t)} \cdot -f(t) \\ &= -\frac{1}{s(t)} \cdot -(-\frac{d}{dt}S(t)) \\ &= -\frac{d[log(s(t)]}{dt} \quad \square \end{aligned} \tag{5}$$

$$\begin{aligned} \lambda(t) &= -\frac{d[log(s(t)]}{dt} \\ \int_0^t \lambda(u)du &= \int_0^t -\frac{d}{du}[log(s(u)]du \\ \Lambda(t) &= -log(S(t)) \end{aligned} \tag{6}$$

$$S(t) = exp(-\Lambda(t)) \quad \square \tag{7}$$

Besides, Aalen (1978) showed that the T and Δ can be seen as generated by a counting process. Using the Doob-Meier decomposition we can then split the process into a compensator part, used as a model, and a martingale part, used as an error process with mean zero and independent increments. For martingales a central limit theorem exists which provides the basis for the asymptotic properties of the model estimates (Andersen et al., 1985). Therefore, let $N(t) = I(T \leq t, \Delta)$ be an n-dimensional counting process with conditional intensity:

$$\begin{aligned} \lambda(t|L) &= R(t)L^T\beta(t) \\ &= R(t)[\beta_0(t) + L_1\beta_1(t) + L_2\beta_2(t) + \cdots + L_p\beta_p(t)] \end{aligned} \tag{8}$$

In which $R_i(t) = I(t \leq T_i)$ is a n-dimensional binary vector describing if individuals are still at-risk for the event at time t. And $\beta(t)$ is the vector containing $p+1$ time-varying regression functions. The regression function $\beta_0(t)$ can be seen as the baseline hazard rate. Whereas the regression functions $\beta_j(t)$ capture the impact of covariate L_j on the hazard rate at time t, with $j = 1, .., p$. It can be shown that the cumulative hazard function $\Lambda(t)$ is then a compensator such that $M(t) = N(t) - \Lambda(t)$ is a martingale. Therefore, we can write the increment form of the additive hazards model as:

$$\begin{aligned} dN(t) &= \lambda(t) + dM(t) \qquad (9) \\ &= R(t)L^T\beta(t) + dM(t). \end{aligned}$$

Due to the martingale property $E[dM(t)] = 0$. For that reason, we can estimate the cumulative regression coefficients $B(t) = \int_0^t \beta(s)ds$ using ordinary least squares. Therefore, we use $D(t) = R(t)L$, with dimensions $(n \times p+1)$, as the design matrix. If an individual i is no longer at-risk for the event at time t, due to earlier death or censoring, than the ith row of the design matrix contains only zeros. Whereas, if an individual i is still at-risk, then row i of $D(t)$ contains the covariate vector $L^i(t)^T = [1, L_1^i, L_2^i, \dots, L_p^i]$. After ordering the observed event times $T_1 < T_2 < \dots$, the estimator proposed by Aalen (1989) can be computed as:

$$\hat{B}(t) = \sum_{T_i < t} [D(T_i)^T D(T_i)]^{-1} D(T_i)^T I(T_i).$$

In which $I(t)$ is an $n \times 1$ vector of zeros except for the element corresponding to the individual experiencing the event at time T_i which is set to one. It is likely that some of the regression functions are approximately constant. In that case the fully non-parametric model is unnecessarily complicated to report and does not provide the best bias-variance trade-off. Therefore, we also used the more general semi-parametric model introduced by McKeague and Sasieni (1994) in which some of the regression functions are restricted to a constant:

$$\lambda(t) = R(t)[L^T \beta(t) + Q^T \phi]$$

With Q a $q \times 1$ covariate vector for which time invariant coefficients, ϕ, are estimated. In the analysis first a completely non-parametric model is estimated, and covariates are incrementally removed from L and added to Q. In the special case that all covariates are in Q the model reduces to a Lin Ying additive model Lin and Ying (1994).

To aid interpretation, the estimated cumulative regression functions are individually plotted against time. The slope of the functions shows the overall impact of the covariates on the hazard rate. As described in the previous section, the impact of the regression functions is translated to the survival scale. For this, we calculate the estimated cumulative hazard rate as: $\hat{\Lambda}(t) = \hat{B}(t)^T L$. With corresponding survival function $\hat{S}(t) = exp(-\hat{\Lambda})$, as derived in equation 7. It should, however, be noted that due to the additive structure of the model the survival curves show non-natural behaviour in which the survival curve is not always declining. Overall, these violations are minor. Yet, we only consider general trends in $\hat{B}(t)$ and $\hat{S}(t)$, rather than attempting to interpret all minor changes in the slope. In addition to the survival function the average lifetime is used in the cost-analysis. The average lifetime can be derived using integration-by-parts and the relation $f(t) = -\frac{d}{dt}S(t)$ from equation 4 such that:

$$\begin{aligned} E[T] &= \int_0^\infty t f(t)dt \\ &= -\int_0^\infty t\frac{d}{dt}(t)dt \\ &= -tS(t)|_0^\infty + \int_0^\infty S(t)d(t) \\ &= \int_0^\infty S(t)d(t) \quad \square. \end{aligned} \tag{10}$$

It is, however, necessary to make tail corrections to the result in equation 10. In this research the upper bound of the integral is replaced with the maximum follow-up time as suggested by Efron (1967). Additionally, the individuals who are censored at the maximum follow-up time are considered to have the event, such that the survival curve drops to zero. Although the estimator is consistent, we should keep in mind that it is not necessarily unbiased. Earlier, Zhong and Hess (2009) showed that depending on the underlying distribution and level of censoring the estimate can be biased downward.

Collapsibility

In order to obtain consistent estimates for causal inference it is not only important that the variable of interest is exogenous, but also that the model is collapsible with regard to the omitted variables. The importance of collapsibility in this context was first addressed by Clogg et al. (1992) and further elaborated and defined by Greenland et al. (1999) as discussed below.

Let g be a generalized linear model for the regression of outcome y on vectors x and u. Let equation 11 be the conditional model and equation 12 be the marginal model:

$$g(E[y|x,u]) = \alpha_0 + \alpha_1 x + \alpha_2 u. \tag{11}$$

$$g(E[y|x]) = \tilde{\alpha_0} + \tilde{\alpha_1} x. \tag{12}$$

The regression is said to be collapsible over u, if $\alpha_1 = \tilde{\alpha_1}$. In that case, the coefficient α_1 is not affected by omitting u, as it stays the same irrespective of whether a marginal or conditional model is estimated.

Not all measures of association are collapsible. Most importantly for this thesis, Greenland (1996) and Struthers and Kalbfleisch (1986) showed that the hazard ratio is generally non-collapsible. This makes the cox proportional hazards model only suitable for causal inference under restrictive conditions such that the differences between the estimated coefficients of the marginal and conditional model are limited. In contrast to the proportional hazards model, the additive hazards model is collapsible. In fact, previous authors have published several derivations, among which Tchetgen et al. (2015), Aalen (1989), Aalen (1993), Sjölander et al. (2016), and Martinussen and Vansteelandt (2013). Here we use a simpler derivation from Lergenmuller (2017) which does not require a derivative of the log Laplace transform. Let the additive hazards model be equation 13 and it's cumulative variant be equation 14.

$$\lambda(t|L,U) = \lambda_0 + \beta_l(t)L + \beta_u(t)U. \tag{13}$$

$$\Lambda(t|L,U) = \Lambda_0 + B_l(t)L + B_u(t)U. \tag{14}$$

Then, $\beta_l(t)$ is the difference in hazard resulting from a unit increase in L. That is:

$$\beta_l(t) = \lambda(t|L = l+1, U) - \lambda(t|L = l, U).$$

Prior to showing the derivation for the marginal model $\lambda(t|L = l)$. We use the law of total probability to derive the following intermediate result:

$$\begin{aligned} S(t|l) &= P(T > t|L) \\ &= E_u[P(T > t|L,U)] \\ &= E_u[S(t|L,U)]. \end{aligned} \tag{15}$$

Then, after using equations 3 and 4, equation 15 can be substituted for $S(t|L)$ and equation 14 for $\Lambda(T|L)$. This results in:

$$
\begin{aligned}
\lambda(t|L) =& \frac{f(t|L)}{S(t|L)} \\
=& \frac{-\frac{\partial}{\partial t} S(t|L)}{S(t|L)} \\
=& \frac{-\frac{\partial}{\partial t} E_u[S(t|L,U)]}{E_u[S(t|L,U)]} \\
=& \frac{-\frac{\partial}{\partial t} E_u[exp(-\Lambda(t|L)]}{E_u[exp(-\Lambda(t|L))]} \\
=& \frac{-\frac{\partial}{\partial t} E_u[exp(-\Lambda_0(t|L) - B_l(t)L - B_u(t)U)]}{E_u[exp(-\Lambda_0(t|L) - B_l(t)L - B_u(t)U)]} \\
=& \frac{-\frac{\partial}{\partial t} exp(-\Lambda_0(t|L) - B_l(t)L) \quad E_u[exp(-B_u(t)U)]}{exp(-\Lambda_0(t|L) - B_l(t)L) \quad E_u[exp(-B_u(t)U)]} \\
=& \frac{-(-\lambda_0(t|L) - \beta(t)L)\cancel{exp(-\Lambda_0(t|L) - B_l(t)L)} \quad \cancel{E_u[exp(-B_u(t)U)]}}{\cancel{exp(-\lambda_0(t|l) - B_l(t)L)} \quad \cancel{E_u[exp(-B_u(t)U)]}} + \\
& \frac{\cancel{exp(-\lambda_0(t|l) - B_l(t)L)} \quad - E_u[-\beta(t)U exp(-B_u(t)U)]}{\cancel{exp(-\lambda_0(t|l) - B_l(t)L)} \quad E_u[exp(B_u(t)U)]} \\
=& \lambda_0(t|l) + \beta(t)L + \frac{E_u[\beta(t)U exp(-B_u(t)U)]}{E_u[exp(B_u(t)U)]} \\
=& \tilde{\lambda_0}(t|l) + \beta(t)L. \qquad (16)
\end{aligned}
$$

In which $\tilde{\lambda_0}(t|l) = \lambda_0(t|l) + \frac{E_u[\beta(t)U exp(-B_u(t)U)]}{E_u[exp(B_u(t)U)]}$.

Result 16 shows that conditioning only on L, instead of conditioning on both L and U, leads to the same coefficient for L. This coefficient can be obtained from the difference in the marginal hazards, which differs only in terms of a unit increase in L: $\beta_l(t) = \lambda(t|L = l + 1) - \lambda(t|L = l)$.

4.3 Causal graph

In general, to estimate the causal effect of a treatment on an outcome, modelling of the joint distribution of the variables is the starting point. However, if we make no assumption on the relationships between the variables, the factorization of the joint probability distribution becomes quickly infeasible. Due to the combinatorial explosion, the number of parameters that needs to be estimated for the factorization grows exponentially with the number of variables, and the values these variables can take. Therefore, it is necessary to describe the relationships between variables and make conditional independence assumptions. Important earlier work on the d-separation theorem proved the equivalence between independence in a graph and independence in a distribution, which helps deciding which variables to include in the model and which variables to leave out. Therefore, we use causal graphs as a graphical tool to explicitize and communicate these assumptions. In causal graphs variables are represented by nodes which are connected by directed edges if one is causing the other. For cause and effect to be clearly distinct, causal graphs cannot be cyclical. Furthermore, causal information can only flow in the direction of the arrows, whereas association can flow in either direction. Conditioning on a variable can block an associational path but also open it, depending on the arrangement of the nodes. More specifically, three fundamental arrangements can be distinguished. 1) Chain: $A \rightarrow B \rightarrow C$, in which conditioning on node B can block the association between nodes A and C. 2) Fork: $A \leftarrow B \rightarrow C$. In which node B is a common cause of node A and C. In the fork arrangement, nodes A and C are correlated, even though node A does not affect node C or vice versa. For this reason, node B is called a confounder as it mixes up the relationship between nodes A and C. Conditioning on node B blocks the associational path between nodes A and C. 3) Collider: $A \rightarrow B \leftarrow C$. In this arrangement nodes A and C share a common effect B on which they are colliding, therefore node B is called a collider. Nodes A and C are not associated, except for if one conditions on node B. In fact, conditioning on a variable that is directly or indirectly affected by a collider also opens the associational path between the common causes of that collider. Of the three arrangements only the chain arrangement can transfer causal association. The others transfer non-causal association and pollute the measurement.

The causal effect of X on Y is identified if it can be expressed in terms of statistical quantities that can be observed. Do-calculus provides the complete set of the most liberal conditions that need to be met for the causal effect to be identified. However, in this research the sufficient but more stringent backdoor criterion suffices (Pearl, 1993). The backdoor criterion states that: A set of variables V satisfies the backdoor criterion relative to X and Y in a graph if the following are true.

1) V blocks all paths between X to Y that contain an arrow into X.

2) V does not contain any nodes of which X is a direct or indirect cause.

In case the conditioning set V satisfies the backdoor criterion than the association between X and Y given V is the causal effect. For a more detailed discussion on do-calculus and identification we refer the reader to pp. 85–86 of the book Causality by Pearl (2009).

The causal graph used for the current research is presented in Figure 1 which incorporates the relevant variables of the database. The graph is constructed based on expert opinion and the medical literature regarding confounders and selection mechanisms for waiting time, waiting list dropout, and survival after LT for HCC. Variables that share similar arrangements and that cover similar information are grouped for clarity. A more detailed version of the causal graph is provided in Appendix C. In the causal graph three major parts can be distinguished focusing on: (1) Waiting time, (2) dropout, and (3) survival.

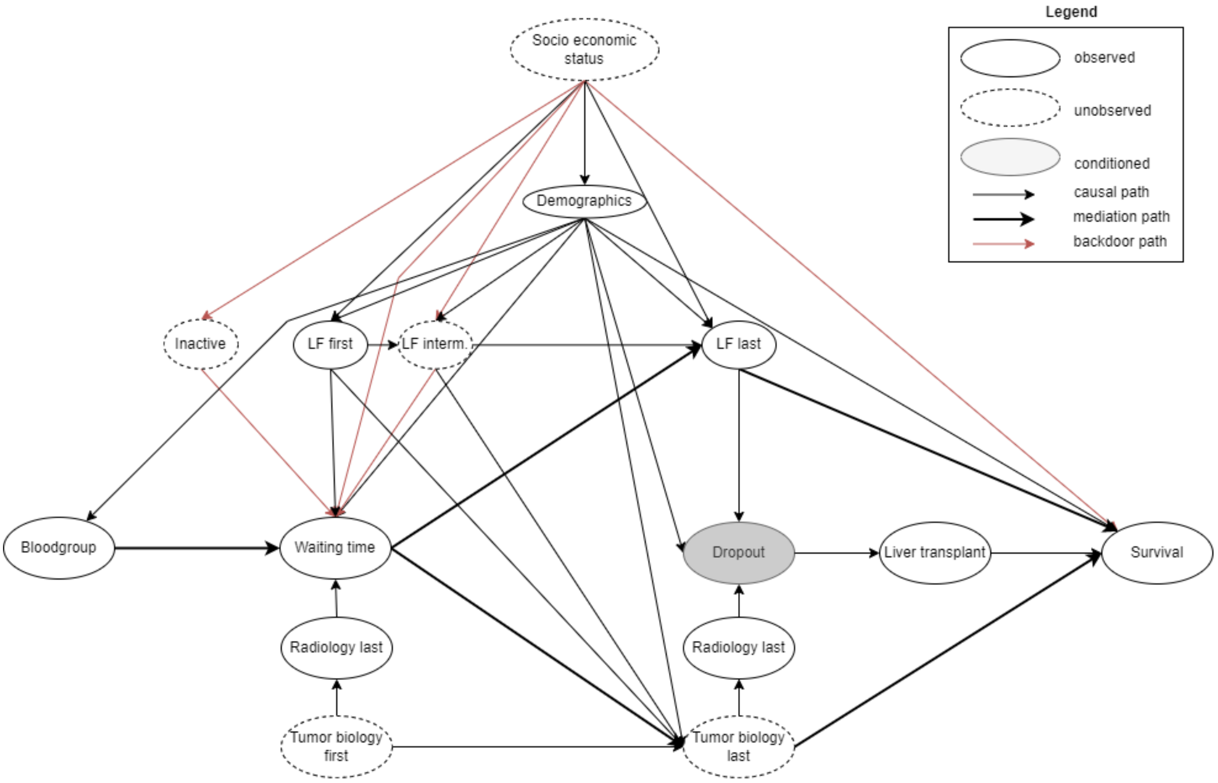

Figure 1: Causal graph displaying the relationships between the variables.

The first part is concerned with the waiting time and is based on the allocation by-laws of the OPTN HRSA/OPTN (2012). These state that the rank of a patient in the waiting list is determined by the liver function at listing (as measured by the MELD score), but also their liver function throughout the waiting period. Additionally, all recipients and donors need to be matched with regard to their blood groups. Lastly, waiting time is affected by a range of situations in which there is no contact with the candidate, the candidate wants to wait longer, there are insurance issues, there is medical non-compliance, there is substance abuse, or the patient is temporarily too sick to undergo surgery. In the graph these situations are summarized in the node *inactive*. In turn, inactivity and liver function are affected by socioeconomic status (e.g., Level of education, income, and type of insurance) and demographic characteristics (e.g., gender, age, ethnicity, and comorbidities). In addition, socioeconomic status directly influences the waiting time by affecting the frequency of hospital visits. The frequency of hospital visits for patients with low economic status is generally lower due to, for example, their insurance policy, ability to arrange a transport or leave of absence. The frequency of hospital visits is directly related to how long it takes for a deterioration of the patient's liver function to be noticed, and thus related to the delay before a patient's urgency is escalated on the waiting list. In the causal graph, the second part that can be distinguished concerns the dropout from the waiting list due to ineligibility resulting from tumor progression or death caused by end-stage liver disease. Important to note is that in-eligibility as a consequence of tumor progression is only limited to tumor progression that is measured upon radiological examination (e.g., tumor size, tumor number, macrovascular invasion, and extra-hepatic spread). As these characteristics are observed, these are distinct from the tumor biology which encapsulates the information on the tumor progression that remains unobserved (e.g., microvascular invasion, micro metastasis, differentiation, and genetic sub-types). The third and last part in the graph describes the causes of survival and focuses only on those elements that are related to waiting time and dropout. These are: the last liver function, the transplant itself, the tumor biology, patient demographics, and the socioeconomic status. Other causes of death that are only related to liver transplantation, but not the waiting time, (e.g., infection or graft failure) are only shown in the detailed version in Appendix C.

From the graph we can observe that the relationship between waiting time and survival is biased by multiple mechanisms. More specifically, the dropout from the waiting list induces selection bias, as only those without the most aggressive sub-types and good liver function make it to the liver transplantation. In the analysis of post-transplant survival, by definition, only patients that received the transplantation are analyzed. Hereby we necessarily condition on the variable dropout depicted in the graph by the shading in gray. Furthermore, we can observe that socioeconomic status is a common cause of waiting time and survival. The variables *inactivity*, *intermediate liver function* and *socioeconomic status* are not observed. Therefore, these paths can not be blocked by means of conditioning and a confounding bias remains. In the causal graph these backdoor paths are highlighted in red. As in this thesis we attempt to isolate the causal effect we will address selection bias first, whereafter we will discuss instrumental variable analysis to account for confounding without the necessity to perform a randomized controlled trial.

4.4 Selection bias - Inverse probability weighting

Selection bias concerns selection of patients to enter the analysis, whereas confounding concerns selection of patients to enter the treatment. For confounding, random assignment of the treatment can eliminate the bias, however, it leaves the selection bias in place. We will adjust for this latter type of bias by means of inverse probability weighting (IPW). With IPW the probability that patients are censored from the analysis is estimated using logistic regression. Using this model, all patients receive a weight that is the inverse of the predicted probability that they are dropped from the waiting list and censored from the analysis. In this way, the patients that were more likely to dropout are weighted heavier. They represent not only themselves but also others like them that did not make it into the analysis. For IPW to account for all selection bias two assumptions are important. First, it is assumed that patients that dropped out versus patients that received the transplantation are exchangeable within the levels of a conditioning set. In other words, all paths with a common cause between dropout and survival are blocked. Secondly, it is assumed that all probabilities of the levels of the conditioning set, given whether a patient is censored or not, are larger than zero. In the causal graph presented in Figure 1 these assumptions are met. The specification of the IPW model used in this thesis is specified in equation 17.

$$\begin{aligned} logit(dropout) =& \psi_0 Gender + \psi_1 Age + \psi_2 Age^2 + \psi_3 BMI + \psi_4 BMI^2 + \psi_5 Functional\ status + \\ & \psi_6 Life\ support + \psi_7 Educ + \psi_8 Ethn + \psi_9 Payment\ type + \psi_{10} Region + \\ & \psi_{11} ALBI^{lt} + \psi_{12} ALBI^{lt\,2} + \psi_{13} MELD^{lt} + \psi_{14} MELD^{lt\,2} + \psi_{15} Enceph + \psi_{16} Ascites + \\ & \psi_{17} Cirrhosis + \psi_{18} Tum\ num^{lt} + \psi_{19} Tum\ size^{lt} + \psi_{20} Tum\ size^{lt\,2} + \\ & \psi_{21} Tum\ size\ tot^{lt} + \psi_{22} Tum\ size\ tot^2 + \psi_{23} log_{10}(AFP^{lt}) \end{aligned} \tag{17}$$

After estimation, the weight for all transplanted patients is calculated as:

$$weight = \frac{1}{P[d = 0|demographics, LF\ last, Radiology]}.$$

In which d is a binary variable indicating whether a patient dropped out (d=1) or not (d=0). The other terms form the conditioning set to ensure exchangeability and are inline with the causal graph presented in Figure 1.

4.5 Confounding bias - Instrumental variables

Confounding bias concerns backdoor paths that contain hidden confounders U which can not be blocked by conditioning on observed variables. If such paths exist, then the measured association between a covariate X and the outcome Y will be a mix of causal and non-causal association. Or in terms of regression, the X will be endogenous, as the unobserved confounder U will be absorbed into the error term, hereby inducing a correlation between the covariate X and the error term. As a result, the regression coefficients will be inconsistent. To avoid this bias, and obtain consistent estimates, there needs to be a source of exogenous variation, occurring either naturally or as a result of experimentation. If such a variable, also known as an instrument Z, is measured, then IV can be used. For IV to provide consistent estimates it is, however, necessary that the following IV conditions are satisfied:

C1) Independence - Z is randomly assigned given covariates L.
C2) Relevance - Z is correlated with X
C3) Exclusive - Z only effects Y through X

These three IV conditions can also be displayed in the causal graph presented in Figure 2. In this diagram the first condition is represented by the absence of an arrow between Z and U, and the blocked path $Z \leftarrow L \rightarrow Y$. Hereby depicting that the value of Z is determined in isolation of the other variables acting on Y. Condition 2 is represented in the diagram by the arrow between Z and X, indicating that Z changes the value of X. Lastly, condition 3 is depicted as the absence of a direct arrow between Z and Y.

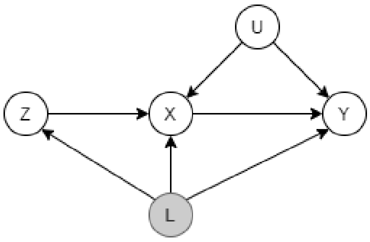

Figure 2: Causal graph for instrumental variable analysis.

From the graph it can be seen why the IV analysis avoids confounding biases. Namely, the flow of association on path $Z \rightarrow X \leftarrow U \rightarrow Y$ is blocked, as X is a collider. Furthermore, the path $Z \leftarrow L \rightarrow Y$ is blocked due to the conditioning on L. Therefore, the only open path for the association to flow from Z to Y is the path $Z \rightarrow X \rightarrow Y$. However, as changes in Z often cause less than a unit change in X the causal association between Z and Y needs to be scaled by the reciprocal of the association between Z and X. The association between Z and X is causal, as there is no hidden variable causing both Z and X. The result is an estimate of the causal association between X and Y without measuring the potential hidden confounders U. Following Heckman (2000), the essence of IV can be summarized as:

$$\beta_{IV} = \frac{dy/dz}{dz/dx}. \tag{18}$$

What remains is to find consistent estimates for the derivatives in the numerator and denominator. Within this framework, Wald (1940) was the first to develop an estimator for binary instruments and a binary treatment. Later, Theil (1953) and Basmann (1957) provided a generalisation with two-stage least squares (2SLS). In which in the first stage, the instrument Z is regressed on covariate X, whereafter Y is regressed on the predicted values of the first stage $\hat{X}$. In this generalized version, multiple, real, and discrete valued instruments could be included in the first stage, hereby expanding the use cases and aiding the efficiency of the estimator. Furthermore, control variables L can be added to both stages to correct for violations of the IV conditions. More involved, however, is the estimation of the standard error which needs to account for the uncertainty in both the first and second stage. In this thesis we will make use of 100 bootstrap samples to estimate the standard error. Furthermore, it should be noted that, although the IV-estimator is consistent, it is not necessarily unbiased in a finite sample. Therefore, the analysis is most suited to large datasets in which the instrument Z is sufficiently correlated with the treatment X. When observing the IV estimator in equation 18, it is clear that the weaker the association between Z and X, the smaller the derivative dz/dx, and the larger the inflation factor has to be. However, small violations of the IV conditions and extremes due to sampling variation are scaled-up by the same inflation factor, potentially introducing a larger bias compared to the confounding bias we try to avoid. Because for the 2SLS estimator the same argument holds, it is important to check if the coefficients of the instruments in the first stage are not too close to zero. This will be assessed using the F-statistic of the first stage for both the full population as well as any population subsets. By rule of thumb the F-statistic needs to be larger than 10, in that case small sample bias is considered negligible (Staiger and Stock, 1994).

In addition to the IV conditions 1-3, a fourth identifying condition is needed for identification and to ensure correct interpretation of the causal estimate. Several conditions have been suggested. These can be divided in homogeneity conditions and monotonicity conditions. However, because we can only partially test the homogeneity condition as some, but not all, confounders are measured we will rely on the following monotonicity condition. The monotonicity condition for continuous treatment by Angrist et al. (2000) demands that the effect of Z on X is uniform across all individuals. More specifically we will assume:

4) Monotonicity - For any pair of values the instrument can take (z, z') there needs to hold that:

$$\text{Either} \quad P[X_i(Z=z) \geq X_i(Z=z')|L] = 1 \quad \forall i$$

$$\text{Or} \quad P[X_i(Z=z) \leq X_i(Z=z')|L] = 1 \quad \forall i.$$

If satisfied, the estimate can be interpreted as the weighted average of the effect of X_i on Y_i on the level of individual patients. When violated, the weights are no longer guaranteed to sum to one and can be non-negative, rendering the coefficients meaningless (Angrist et al., 2000).

Application

As discussed before, in this thesis the hazard rate $\lambda(t)$ is the outcome measure and, as shown in the causal graph, waiting time W is endogenous due to the existence of unobserved confounding. We will now show that blood group can be used as an instrument by discussing the IV conditions.

C1) Independence - The blood group is in part determined by the parents which each give two of their alleles to the child, of which each allele is either A, B, or O. At random, from each parent, one of the alleles is discarded. The two alleles that remain determine the blood group of the child. Although the allele selection is random, the distribution of A, B, and O is not perfectly uniform throughout the population with slight differences in prevalence across demographic groups. To ensure that no open backdoor paths through demographics exist, ethnicity, $ethn$, is absorbed into the conditioning set L. Imbalances in baseline characteristics will be investigated as these can be suggestive of imbalances across unmeasured variables. For this, the standardized mean difference (SMD) and the Kolmogorov-Smirnov (KS) statistic will be used with thresholds 0.25 and 0.1, respectively (Rubin, 2001).

C2) Relevance - As a mismatch in blood groups between donor and recipient has a high risk of an immune response that rejects the organ, blood group matching is standard practice for liver transplantations (HRSA/OPTN, 2012). However, from which donor blood groups a recipient can receive is asymmetric (Table 2). For example, a recipient with blood group AB can receive a donor liver from someone with blood group AB, A, B, or O, while a recipient with blood group O can only receive a donor liver from a donor with blood group O. This results in that even if a recipient with blood group O is the highest on the waiting list he or she might need to wait longer until a suitable organ is available in comparison to a recipient with blood group AB.

Table 2: Compatible blood groups between receiver and donor

Receiver	**Compatible donor**
O	O
A	A, O
B	B, O
AB	AB, A, B, O

C3) Exclusive - To the best of our knowledge, no studies exist which describe a biological interaction between the blood group and liver cells. In extension, there is no evidence that the blood group is related to the tumor size, number, or recurrence. In addition, given an appropriate matching of donor and recipient, the blood group is not correlated to rejection.

C4) In the context of this thesis and the UNOS-OPTN allocation policy (HRSA/OPTN, 2012), assuming monotonicity is justified, as the waiting times vary between individuals but never increases if a patient were, hypothetically, assigned a more favorable blood group.

Given that the assumptions are justified, we will now connect the estimand to the estimate using the relations introduced in the previous subsection as follows:

$$\begin{aligned}\beta(t) &= \lambda(T=t|do(W=w+1)) - \lambda(T=t|do(W=w))\\ &= E[\lambda_i(T=t|do(W_i=w+1))] - E[\lambda_i(T=t|do(W_i=w))].\end{aligned}$$

Furthermore, let the instrumental variable G represent the blood group with domain $g \in \{ab, b, a, o\}$. Then, using the exclusivity condition C3, we can write for two arbitrary levels (g, g') the following:

$$\beta(t)^{g,g'} = \frac{E[\lambda_i(T=t|W_i(do(G_i=g')))] - E[\lambda_i(T=t|W_i(do(G_i=g)))]}{E[W_i(do(G_i=g'))] - E[W_i(do(G_i=g))]}.$$

By the independence condition C1, we can remove the *do* operators. In addition, due to the relevance condition C2, the denominator can not be zero. Therefore, we can write the estimand as:

$$\beta(t)^{g,g'} = \frac{E[\lambda_i(T=t|W_i(G_i=g'))] - E[\lambda_i(T=t|W_i(G_i=g))]}{E[W_i(G_i=g')] - E[W_i(G_i=g)]}. \tag{19}$$

To aid interpretation we will show that the estimate can be viewed as the weighted average of the patient level treatment effect of W on the hazard difference. For the denominator of equation 19 we can write, using the distributive property of the expectation and monotonicity assumption, the following:

$$\begin{aligned}&E[W_i(G_i=g')] - E[W_i(G_i=g)]\\ =&E[W_i(G_i=g') - W_i(G_i=g)]\\ =&\int_0^\infty P[W_i(G_i=g) < W < W_i(G_i=g')]dW.\end{aligned} \tag{20}$$

For the numerator we have:

$$\begin{aligned}
&E[\lambda_i(T=t|W_i(G_i=g')] - E[\lambda_i(T=t|W_i(G_i=g))] \\
=&E[\lambda_i(T=t|W_i(G_i=g')) - \lambda_i(T=t|W_i(G_i=g))] \\
=&\int_0^\infty E[\frac{\partial\lambda(t|W)}{\partial W}|W_i(G_i=g) < W < W_i(G_i=g')] * \\
&P[W_i(G_i=g) < W < W_i(G_i=g')]dW. \qquad (21)
\end{aligned}$$

Taking the ratio of the denominator 20 and numerator 21 gives:

$$\beta(t)^{g,g'} = \int_0^\infty E[\frac{\partial\lambda(t|W)}{\partial W}|W_i(G_i=g) < W < W_i(G_i=g')]\omega(W)dW. \qquad (22)$$

In which:

$$\omega(W) = \frac{P[W_i(G_i=g) < W < W_i(G_i=g')]}{\int_0^\infty P[W_i(G_i=g) < W < W_i(G_i=g')]dW}$$

Next, we consider all levels of the blood group instrument G which are ordered such that $W(g^k) \leq W(g^{k+1}) \quad \forall\, k \in G$. A combined estimate can then be obtained by averaging the individual comparisons of equation 22. That is:

$$\beta(t) = \sum_{k=1}^{K} \lambda_k \beta^{g^{k-1},g^k} \qquad (23)$$

With:

$$\lambda_k = \frac{E[W_i(g^k) - W_i(g^{k-1})]\sum_{j=k}^{K} f_g(g^j)(E[W_i|g^j] - E[W_i])}{\sum_{m=1}^{K} E[W_i(g^m) - W_i(g^{m-1})]\sum_{j=m}^{K} f_g(g^j)(E[W_i|g^j] - E[W_i])}$$

Therefore, we can interpret the IV estimate as resulting from three levels of weighted averages (I.e., individuals, wait times, and levels of the instrument). The weights for the individuals are uniform, for the wait times they are proportional to how likely wait times are to occur, and for instruments the weights are proportional to the strength of the association between the blood group and waiting time. Furthermore, this Wald type estimator can be generalized using two regressions similar to 2SLS. In this thesis the first stage the relationship between the instrument and the waiting time is modelled and estimated using least squares as follows:

$$E[W_i|G] = E[W_i(G)] = \alpha_0 + \alpha_1 G + \alpha_2 ethn. \qquad (24)$$

In the second stage the earlier introduced additive hazard model is used and specified as:

$$E[\lambda_i(t|W_i(G))] = \beta_0(t) + \beta_1 E[W_i(G)] + \beta_2 ethn. \qquad (25)$$

4.6 Heterogeneous treatment effect

In this section we discuss the techniques used to investigate if some patients tend to trade-in more post-transplant survival by waiting longer compared to others. To address this question, first we evaluated and contrasted the effect of waiting between the first and third quartile for each of the variables measured at listing; tumor number, tumor size, $log_{10}(AFP)$, and ALBI score. Also, the associated survival difference between 2 and 12 months of waiting was inspected. For ease of reporting all regressions were performed assuming a time constant effect. In addition to this descriptive analysis, we used interaction terms in the regression to investigate effect moderation. Because the interaction term of an exogenous moderator and an endogenous waiting time (W) is endogenous, we used the interaction of the exogenous moderator and exogenous blood group (G) as an instrument. In addition to more instrumental variables each interaction was estimated in a dedicated first stage regression. Let Z be the collection of instruments, the total first stage then becomes:

$$\begin{aligned}
E[W_i|Z] &= \alpha_{1,0} + \alpha_{1,1}G + \alpha_{1,2}G * Tum\ num^{list} + \alpha_{1,3}Tum\ num^{list} + \\
&\quad \alpha_{1,4}G * Tum\ size^{list} + \alpha_{1,5}Tum\ size^{list} + \alpha_{1,6}G * AFP^{list} + \\
&\quad \alpha_{1,7}AFP^{list} + \alpha_{1,8}G * ALBI^{list} + \alpha_{1,9}ALBI^{list} + \alpha_{1,10}Ethn \\
E[W_i * Tum\ num^{list}|Z] &= \alpha_{2,0} + \alpha_{2,1}G + \alpha_{2,2}G * Tum\ num^{list} + \alpha_{2,3}Tum\ num^{list} + \\
&\quad \alpha_{2,4}G * Tum\ size^{list} + \alpha_{2,5}Tum\ size^{list} + \alpha_{2,6}G * AFP^{list} + \\
&\quad \alpha_{2,7}AFP^{list} + \alpha_{2,8}G * ALBI^{list} + \alpha_{2,9}ALBI^{list} + \alpha_{2,10}Ethn \\
E[W_i * Tum\ size^{list}|Z] &= \alpha_{3,0} + \alpha_{3,1}G + \alpha_{3,2}G * Tum\ num^{list} + \alpha_{3,3}Tum\ num^{list} + \\
&\quad \alpha_{3,4}G * Tum\ size^{list} + \alpha_{3,5}Tum\ size^{list} + \alpha_{3,6}G * AFP^{list} + \\
&\quad \alpha_{3,7}AFP^{list} + \alpha_{3,8}G * ALBI^{list} + \alpha_{3,9}ALBI^{list} + \alpha_{3,10}Ethn \\
E[W_i * AFP^{list}|Z] &= \alpha_{4,0} + \alpha_{4,1}G + \alpha_{4,2}G * Tum\ num^{list} + \alpha_{4,3}Tum\ num^{list} + \\
&\quad \alpha_{4,4}G * Tum\ size^{list} + \alpha_{4,5}Tum\ size^{list} + \alpha_{4,6}G * AFP^{list} + \\
&\quad \alpha_{4,7}AFP^{list} + \alpha_{4,8}G * ALBI^{list} + \alpha_{4,9}ALBI^{list} + \alpha_{4,10}Ethn \\
E[W_i * ALBI^{list}|Z] &= \alpha_{5,0} + \alpha_{5,1}G + \alpha_{5,2}G * Tum\ num^{list} + \alpha_{5,3}Tum\ num^{list} + \\
&\quad \alpha_{5,4}G * Tum\ size^{list} + \alpha_{5,5}Tum\ size^{list} + \alpha_{5,6}G * AFP^{list} + \\
&\quad \alpha_{5,7}AFP^{list} + \alpha_{5,8}G * ALBI^{list} + \alpha_{5,9}ALBI^{list} + \alpha_{5,10}Ethn. \qquad (26)
\end{aligned}$$

With corresponding second stage:

$$\begin{aligned}
E[\lambda_i(t|Z)] =& \beta_0(t) + \beta_1 E[W_i|Z] + \beta_2 E[W_i * Tum\ num^{list}|Z] + \beta_3 Tum\ num^{list} + \\
& \beta_4 E[W_i * Tum\ size^{list}|Z] + \beta_5 Tum\ size^{list} + \beta_6 E[W_i * AFP^{list}|Z] + \\
& \beta_7 AFP^{list} + beta_8 E[W_i * ALBI^{list}|Z] + \beta_9 ALBI^{list} + \beta_{10} Ethn. \qquad (27)
\end{aligned}$$

The structural multicollinearity between the main effect and the interaction terms of the moderators was resolved by centering each of the variables.

4.7 Cost analasysis of the Test-of-time

The rationale of the test-of-time is based on that the population of patients consists of a mixture of two subtypes of HCC: a subset of patients with more a aggressive disease, and a subset with a more indolent disease. From a utilitarian perspective it is then desirable to filter out patients with the more aggressive type of cancer. Both to increase the average post-transplant survival and to prevent a risky operation for those experiencing minimal benefit from the transplantation. In the proposed test-of-time, patients are observed over a period to characterise the disease and to filter out the aggressive cases. However, there exists a trade-off, as also the patients with a more indolent type of cancer deteriorate. It remains unclear if the gain from the better allocation of a fraction of the livers offsets the reduction in survival due to the extra waiting. Here, we will investigate under what conditions the test-of-time results in an increase in the average lifetime. To structure the discussion, let A be a binary random variable with value 1 for a patient with an aggressive cancer, and 0 otherwise. Let p_1 be the proportion of aggressive cases that are transplanted without the test-of-time. And let p_2 be the proportion of aggressive cases transplanted with the test-of-time. Further, it holds that $p_2 = p_1(1 - eff)$, with eff being the percentual reduction that characterises the efficiency at which the test-of-time purifies the mixture. Furthermore, let T be the random variable indicating the time-to-event, inline with the earlier notation. Lastly, let c be the cost of waiting in years of post-transplant survival. We can then write the average lifetime without the test-of-time as follows:

$$\begin{aligned}
&E[T|A=1]\,p_1 + E[T|A=0]\,(1-p_1)\\
=&E[T|A=1]\,p_1 + E[T|A=0] - E[T|A=0]\,p_1\\
=&E[T|A=1]\,p_1 - E[T|A=0]\,p_1 + E[T|A=0].
\end{aligned}$$

And the average lifetime with the test-of-time as:

$$\begin{aligned}
&(E[T|A=1] - c\,w)\,p_2 + (E[T|A=0] - c\,w)(1-p_2)\\
=&E[T|A=1]\,p_2 - c\,w\,p_2 + E[T|A=0] - c\,w - E[T|A=0]\,p_2 + c\,w\,p_2\\
=&E[T|A=1]\,p_2 - \cancel{c\,w\,p_2} + E[T|A=0] - c\,w - E[T|A=0]\,p_2 + \cancel{c\,w\,p_2}\\
=&E[T|A=1]\,p_2 - E[T|A=0]\,p_2 - c\,w + E[T|A=0] - c\,w.
\end{aligned}$$

Then a benefit is realized if:

$$\begin{aligned}
&E[T|A=1]\,p_2 - E[T|A=0]\,p_2 - c\,w + E[T|A=0] > E[T|A=1]\,p_1 - E[T|A=0]\,p_1 + E[T|A=0]\\
&E[T|A=1]\,p_2 - E[T|A=0]\,p_2 - c\,w + \cancel{E[T|A=0]} > E[T|A=1]\,p_1 - E[T|A=0]\,p_1 + \cancel{E[T|A=0]}\\
&(E[T|A=1] - E[T|A=0])\,p_2 - c\,w > (E[T|A=1] - E[T|A=0])\,p_1\\
&-c\,w > -(E[T|A=1] - E[T|A=0])\,p_2 + (E[T|A=1] - E[T|A=0])\,p_1\\
&E[c]\,w < (E[T|A=1] - E[T|A=0])(p_2 - p_1)\\
&w < \frac{(E[T|A=1] - E[T|A=0])(p_2 - p_1)}{c}\\
&w < \frac{(E[T|A=1] - E[T|A=0])(p_1(1-eff) - p_1)}{c}\\
&w < \frac{(E[T|A=1] - E[T|A=0])(\cancel{p_1} - p_1\,eff - \cancel{p_1})}{c}\\
&w < \frac{(E[T|A=0] - (E[T|A=1])(p_1\,eff)}{c}
\end{aligned} \tag{28}$$

The inequality in expression 28 gives the upper bound of the duration of the test-of-time. From this expression we can observe that four elements are important. 1) the difference in average time-to-event between aggressive and non-aggressive cases, 2) the proportion of aggressive cases without the test-of-time, 3) the efficiency of the test of time, and 4) the cost of waiting. Only the cost of waiting was studied and quantified in this research. For the other components we used credible ranges to calculate the upper bounds. More specifically, for the difference in average time-to-event we used a range of 7 to 16 years; for the proportion of aggressive cases without the test-of-time we used a range of 2 to 30 percent; and lastly, we used a range of 10 to 70 percent for the effectiveness of the test-of-time.

5 Results

5.1 Baseline characteristics

Table 3 shows the descriptive statistics of the patients listed for transplantation with HCC, stratified per blood group. The data is unweighted, and the variables were measured at listing. The table shows, as discussed in the previous section, that blood groups are not uniformly distributed over the ethnicities. These are therefore absorbed into the conditioning set to ensure unconfoundedness of the instrument. Furthermore, we observe that waiting time is different across the blood groups, and as a result so are the dropout rate and post transplant survival. All other covariates, in particular those that strongly affect post-transplant survival (E.g., ALBI score, tumor number and tumor size) are balanced at time of listing. The standardized mean differences are all below the threshold of 0.25 and the maximum KS statistics are below 0.1. A more detailed overview of the covariate balance statistics for each blood group contrast is presented in Appendix D.

Table 3: Descriptive statistics at listing

		AB	B	A	O	SMD (max)	KS (max)
n		1797	5877	16818	21202		
Period n (%)	[2000,2005]	268 (15)	966 (16)	2642 (16)	3391 (16)	0.0418	0.0152
	(2005,2010]	458 (25)	1435 (24)	4119 (24)	5258 (25)	0.0248	0.0107
	(2010,2015]	596 (33)	1926 (33)	5485 (33)	6872 (32)	0.0161	0.0075
	(2015,2020]	475 (26)	1550 (26)	4572 (27)	5681 (27)	0.0183	0.0081
Male	Missing	0 (0)	0 (0)	0 (0)	0 (0)	-	-
	n (%)	1404 (78)	4513 (77)	13061 (78)	16165 (76)	0.0450	0.0189
Age	Missing (%)	0 (0)	0 (0)	0 (0)	0 (0)	-	-
	Q1 | Q2 | Q3	55 | 60 | 64	54 | 59 | 64	54 | 59 | 64	54 | 59 | 64	0.0655	0.0372
Height (m)	Missing (%)	3 (0)	29 (0)	59 (0)	58 (0)	0.0577	0.003
	Q1 | Q2 | Q3	165 | 173 | 180	165 | 173 | 178	167 | 173 | 180	165 | 173 | 178	0.1092	0.0458
Weight (kg)	Missing (%)	3 (0)	30 (1)	64 (0)	63 (0)	0.0591	0.0034
	Q1 | Q2 | Q3	72 | 84 | 97	70 | 82 | 96	73 | 85 | 98	73 | 84 | 97	0.1478	0.0248
Ethnicity n (%)	Missing (%)	0 (0)	0 (0)	0 (0)	0 (0)	-	-
	White	1163 (65)	3157 (54)	12105 (72)	12908 (61)	0.3814	0.1826
	Black	175 (10)	840 (14)	1043 (6)	2149 (10)	0.2698	0.0809
	Hispanic	146 (8)	772 (13)	2398 (14)	4359 (21)	0.3611	0.1243
	Asian	292 (16)	1045 (18)	1056 (6)	1433 (7)	0.3619	0.115
	Native American	1 (0)	15 (0)	90 (1)	182 (1)	0.1234	0.008
	Pacific Islander	10 (1)	15 (0)	46 (0)	45 (0)	0.0606	0.0034
	Multiracial	10 (1)	33 (1)	80 (0)	126 (1)	0.0161	0.0012
Meld score (listing)	Missing (%)	18 (1)	47 (1)	188 (1)	198 (1)	0.0326	0.0032
	Q1 | Q2 | Q3	8 | 11 | 15	8 | 11 | 15	8 | 11 | 15	8 | 11 | 15	0.0346	0.0254
Albi score (listing)	Missing (%)	13 (0.7)	29 (0.5)	115 (0.7)	119 (0.6)	0.0294	0.0023
	Q1 | Q2 | Q3	-2.4 | -1.9 | -1.4	-2.4 | -1.9 | -1.3	-2.3 | -1.8 | -1.3	-2.3 | -1.8 | -1.3	0.1059	0.032
Tumor number n (%)	Missing	438 (24)	1105 (19)	3126 (19)	3868 (18)	0.1536	0.0613
(listing)	1	1032 (57)	3622 (62)	10337 (61)	13196 (62)	0.0152	0.0176
	2	234 (13)	849 (14)	2449 (15)	3053 (14)		
	3	90 (5)	282 (5)	856 (5)	1024 (5)		
	4	0 (0)	13 (0)	39 (0)	45 (0)		
	>=5	3 (0)	6 (0)	11 (0)	16 (0)		
Tumour size	Missing (%)	438 (24)	1105 (19)	3126 (19)	3868 (18)	0.1536	0.0613
(listing)	Q1 | Q2 | Q3	2 | 2 | 3	2 | 2 | 3	2 | 2 | 3	2 | 2 | 3	0.0221	0.0283
log10(AFP) (ng/mL)	Missing (%)	546 (30.4)	1473 (25.1)	4220 (25.1)	5318 (25.1)	0.1208	0.0532
(listing)	Q1 | Q2 | Q3	0.7 | 1 | 1.6	0.7 | 1 | 1.6	0.7 | 1 | 1.6	0.7 | 1 | 1.6	0.0348	0.0236
Locoregional therapy	Missing	465 (26)	1388 (24)	3905 (23)	5109 (24)	0.0621	0.0266
	n (%)	916 (69)	3278 (73)	9496 (74)	11849 (74)	0.1086	0.0486
Waiting time	Missing (%)	0 (0)	0 (0)	0 (0)	0 (0)	-	-
	Q1 | Q2 | Q3	1 | 3 | 8	2 | 7 | 13	3 | 8 | 15	3 | 8 | 16	0.329	0.2516
Dropout	Missing	22 (1)	126 (2)	438 (3)	594 (3)	0.1078	0.0158
	n (%)	304 (17)	1428 (25)	4470 (27)	6159 (30)	0.2974	0.1276
n		1465	4306	11857	14372		
Med. FU [95%CI]		6 [5.8 - 6.3]	5.9 [5.8 - 6]	5.9 [5.9 - 6]	5.9 [5.8 - 6]		
Death	Missing	0 (0)	0 (0)	0 (0)	0 (0)		
	n (%)	402 (27)	1205 (28)	3512 (30)	4317 (30)		
Med. OS [95%CI]		14.5 [13.5 - NA]	14.3 [12.9 - 15.9]	12.7 [12.3 - 13.3]	12.8 [12.3 - 13.3]		
5 yr OS [95%CI]		0.75 [0.73 - 0.78]	0.75 [0.74 - 0.77]	0.74 [0.73 - 0.75]	0.74 [0.73 - 0.74]		

5.2 Instumental variable conditions

The mean waiting time at each of the levels of the instrument was significantly different with respect to the reference category (AB). In addition, the F-statistic of 86.3 indicated that the blood group as an instrument is sufficiently associated with the waiting time to be used in instrumental variable analysis. Furthermore, the underlying source of variation was confirmed in Table 4. In this table we can observe that for example 291 donor livers with blood group B were assigned to recipients with blood group AB. The opposite, all emergency transplantations, occurred only 3 times. Similarly, 376 donor livers with blood group O were assigned to recipients with blood group B, opposed to just 2 in the opposite direction. The result of the allocation policy can also be observed in the distribution of waiting times shown in Figure 3. Interesting is that the effect of the mandatory 6-month waiting period can also be clearly noticed. At the start there is an over representation of patients with AB and B, where in the tail of the distribution there are more patients with blood groups O and A left.

Table 4: Blood group matrix

		Donor			
		AB	B	A	O
Recipient	AB	999	291	162	13
	B	3	3927	0	376
	A	2	0	11684	171
	O	1	2	233	14136

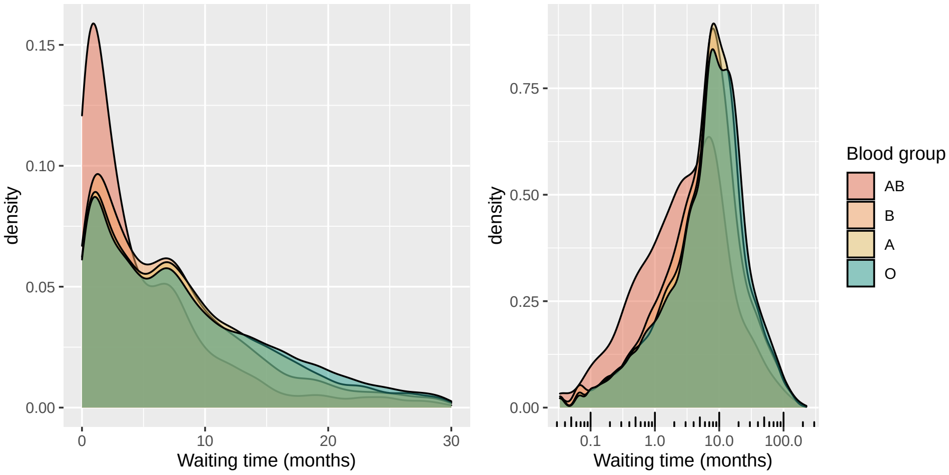

Figure 3: Distribution of waiting time stratified per blood group

Lastly, in an attempt to falsify the monotonicity assumption, the stratified empirical cumulative distribution functions are presented in Figure 4. Although we observe that the lines overlap at the extremes the lines do not cross, hereby indicating that also empirically the monotonicity assumption is justified.

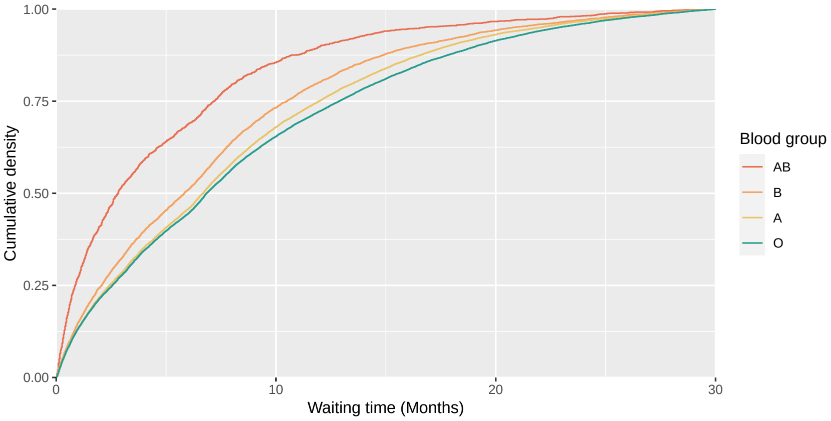

Figure 4: Empirical cumulative density function stratified per blood group

5.3 Average treatment effect

The two stage least squares regression resulted in the cumulative hazard function presented in Figure 5 panel A. In panel B the ATE point estimate is translated to the survival scale and displays the contrast between 2 months and 12 months of waiting. In panel A we observe that the cumulative hazard is increasing and that the zero line is not fully included in the 95% confidence interval, hereby indicating that the hazard rate increases significantly with a longer waiting time. Furthermore, we notice that the model with time constraint coefficient largely overlaps with the time varying model. The time constant effect, $\gamma * 10^{-4}$, was estimated to be 0.45 %CI [0.15; 0.76]. On the survival scale the effect of waiting 12 months versus 2 months corresponded with a drop in overall survival of 5.15% 95%CI [0.36; 9.89] at 5-year post-transplantation and 8.49% 95% CI [0.61; 15.93] at 10 years post-transplantation. The median survival dropped by 3.37 years from 16.21 95%CI [15.98; 16.61] for those waiting 2 months to 12.84 95%CI [10.74; 15.93] for those waiting 12 months.

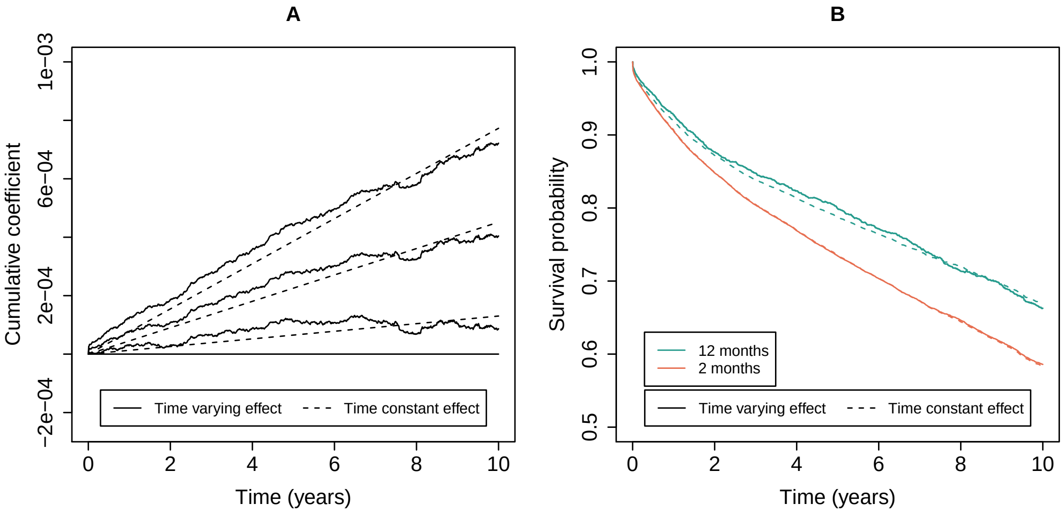

Figure 5: Average treatment effect of waiting time

5.4 Heterogeneous treatment effect

Figure 6 presents the Pearson correlations among the four variables measured at listing that could moderate the harm of waiting on post-transplant survival. It is important to know the correlation among them when sub-setting the population. For example, when subsetting the population on large tumor size, these patients also have more often a high AFP. All correlations were positive and although being relatively small in magnitude they were statistically significant.

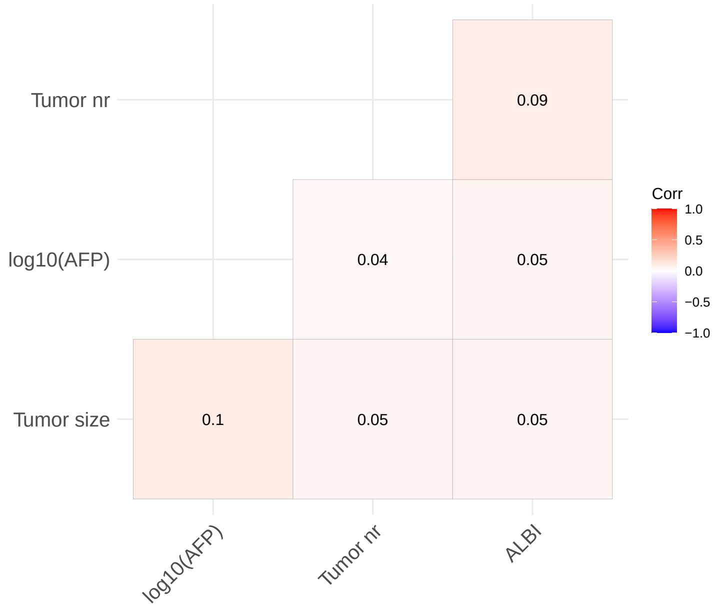

Figure 6: Correlation matrix of potential moderators

Heterogeneity in the treatment effect was investigated on a variable-per-variable basis for which the population was subsetted based on the first and third quartiles. All variables were measured at listing, prior to the waiting period. For each subset the instrumental variable analysis was repeated and displayed in Table 5. The table shows that compared to the estimate for the ATE the confidence intervals have substantially increased due to the reduction in sample size. The estimate for waiting time in the Q3 subgroup of tumor number was the only subset in which the subgroup ATE was significantly different from zero, with $\gamma * 10^{-4}$ of 1.03 95%CI[0.1; 1.79]. Although, in this analysis heterogeneity of the treatment effect was not tested, all estimates indicated a higher impact of waiting on patients that were more sick (Q3) in comparison to patients that were less sick (Q1). Interestingly, the differences in point estimates between Q1 and Q3 are larger for tumor related variables, such as tumor number and size (range of difference in $\gamma * 10^{-4}$: 0.19-0.86), in comparison to the liver function related variable ALBI score.

Table 5: Treatment heterogeneity comparison of quantiles

	Q1				**Q3**			
	level	**F**	$\gamma * 10^{-4}$ **[95%CI]**	**Δ5 yr OS %**	**level**	**F**	$\gamma * 10^{-4}$ **[95%CI]**	**Δ5 yr OS %**
Tumor number	1*	63	0.17 [-0.22; 0.55]	1.9	2	24	1.03 [0.10; 1.97]	14.4
Tumor size	1.8	32	0.35 [-0.34; 1.04]	4.1	2.9	24	0.54 [-0.19; 1.27]	6.1
log10(AFP)	0.7	24	0.28 [-0.32; 0.88]	3.4	1.59	25	0.62[-0.19; 1.44]	6.5
ALBI	-2.37	31	0.39 [-0.16; 0.94]	4.8	-1.33	18	0.43 [-0.37; 1.24]	4.7

Presence of a heterogeneous treatment effect was more formally investigated using interaction terms, with the results presented in Table 6. The interaction terms were all positive, which is in concordance with that sicker patients are harmed more by waiting. However, none of the interaction terms attained statistical significance. The null hypothesis stating that: waiting is equal for all, was therefore not rejected.

Table 6: Heterogenious treatment effect interaction terms

	$\boldsymbol{\gamma}$ * 10^{-4} **[95%CI]**	**p value**
W	0.38 [0.06; 0.82]	0.004
W*tumor number	0.48 [-0.27; 1.20]	0.115
tumor number	70 [26; 114]	<0.001
W*tumor size	0.08 [-0.48; 0.64]	0.938
size	0.12 [-0.55; 0.79]	<0.001
W*log10(AFP)	0.12 [-0.55; 0.79]	0.908
AFP	120 [82; 160]	<0.001
W*ALBI	0.11 [-0.65; 0.87]	0.936
ALBI	81 [20; 140]	<0.001
Ethn. Black	90 [30; 150]	<0.001
Ethn. Hispanic	-130 [-190; -73]	<0.001
Ethn. Asian	-200 [-250; -155]	<0.001
Ethn. Other	-123 [-250; 2.7]	<0.001

5.5 Cost-analysis of the Test-of-time

In Table 7 the cost-analysis of the test-of-time is presented, with on the rows the difference in the average lifetime between aggressive and non-aggressive cases, in years, and on the columns the product term of the proportion of aggressive cases P_1 and the efficiency Eff of the test-of-time. The values of the table represent the upper bound of the duration the test-of-time can take, in months, before doing more harm than good. A maximum duration of the test-of-time of below 4-Months is likely not feasible and, in those conditions, can better be avoided all together. On the contrary, under conditions where the maximum duration is above the 12-months, the policy is likely to improve the average lifetime of those transplanted.

Table 7: Upperbound Test-of-time (Months)

		$P_1 * Eff$									
		0.02	**0.04**	**0.06**	**0.08**	**0.10**	**0.12**	**0.14**	**0.16**	**0.18**	**0.20**
$E[T\|A=0]-E[T\|A=1](Years)$	**7**	1	2	3	4	6	7	8	9	10	11
	8	1	3	4	5	6	8	9	10	11	13
	9	1	3	4	6	7	8	10	11	13	14
	10	2	3	5	6	8	9	11	13	14	16
	11	2	3	5	7	9	10	12	14	16	17
	12	2	4	6	8	9	11	13	15	17	19
	13	2	4	6	8	10	12	14	16	18	20
	14	2	4	7	9	11	13	15	18	20	22
	15	2	5	7	9	12	14	17	19	21	24
	16	3	5	8	10	13	15	18	20	23	25

To further aid interpretation, the product term of the proportion of aggressive cases without the test-of-time and the efficiency of the test-of-time is expanded in Table 8 for the credible ranges earlier discussed in the methodology section.

Table 8: Product term of the proportion of aggressive cases and efficiency level

		P_1					
		0.02	**0.07**	**0.12**	**0.17**	**0.22**	**0.27**
Eff	**0.1**	0.002	0.007	0.012	0.017	0.022	0.027
	0.2	0.004	0.014	0.024	0.034	0.044	0.054
	0.3	0.006	0.021	0.036	0.051	0.066	0.081
	0.4	0.008	0.028	0.048	0.068	0.088	0.108
	0.5	0.010	0.035	0.060	0.085	0.110	0.135
	0.6	0.012	0.042	0.072	0.102	0.132	0.162
	0.7	0.014	0.049	0.084	0.119	0.154	0.189

6 Discussion

The aim of this thesis was to estimate the causal effect of waiting time on post-operative survival. We conclude that for the general population of transplant patients with HCC waiting is harmful. If a patient waited 12 months instead of 2 months, we estimated a loss of 5% overall survival at 5-year after transplantation and a drop of 3.37 years in median survival. The exploratory subgroup analysis suggests that tumor number is the variable that is most likely to moderate the effect of waiting time. However, with the current size of the data set we were not able to demonstrate that patients with a large tumor burden or low liver function were harmed more compared to those less ill. Although our research, by itself, does not advice on liver transplant policies. It does, however, provide essential information that is needed to formulate conditions which policies need to meet in order to result in a net benefit. In this research we performed a cost-analysis for the test-of-time. This brings us closer to answering whether the harm of all patients having to wait longer can be offset by the improved allocation of a fraction of the livers. Based on this thesis, similar conditions can be constructed regarding to what extend patients, for which multiple treatment options are available, should be offered transplantation. Even when LT is the superior treatment, the longer waiting time as a consequence of adding more patients to the waiting list could offset any benefit. Besides quantifying the harm of waiting on post-transplant survival, this thesis provides a comprehensive discussion on the causal inference techniques supporting our results. Furthermore, we provide the first causal graph in the field, which represents our model of reality and simultaneously explicates our assumptions. Another strength of this research is, that due to the instrumental variable analysis our results were subject to minimal confounding bias. In addition, selection bias was addressed using IPW and missing values were imputed using multiple imputation.

Important to note, however, is that our research has some limitations. The first is that the causal graph presented here is our view on reality, and we realize that undoubtedly more detail can be added. In the graph, the most important simplification was that the complexities involved with waiting time being a time varying treatment were simplified to where it is being assigned at once and at listing. We recognize that in reality the waiting time of a patient is not known at the time of listing and that the waiting time depends on intermediate examinations of which the frequency is tailored to individual patients. The simplification is, however, justified by that we do not aim to predict the waiting time, and harm thereof, for an individual patient but rather evaluate it from a policy perspective. Another limitation is that the independence, exclusivity, and monotonicity assumptions can not be empirically validated and remain open to debate. Especially with regard to the exclusivity assumption it is impossible to prove this negative. Although links to ethnicity were earlier described and controlled for in this thesis, we can not guarantee that, until now unknown, genetic patterns act as a confounder between blood group and survival. Wu et al. (2017) and Li et al. (2018) investigated the relationship between blood group and survival in patients with HCC that were treated with resection or transarterial chemoembolization. They found that patients with blood group O had a better prognosis. Their research did, however, not control for ethnicity and socioeconomic factors. Also, accordign to the authors, a biological underlying mechanism is still missing. Furthermore, their findings could not be corroborated by the research of Oral and Sahin (2019). In addition, potential publication bias towards reports that didn't find a significant result hampers interpretation. Regardless, our conclusions would remain valid if the association is present. Namely, the measured harm of waiting would be shrunken by endowing blood group O, that waits the longest, with a survival benefit. In that case our estimates are conservative, and the conclusion that waiting is harmful for a patient would be strengthened. With regards to the monotonicity assumption, we recognize that the assumption might not hold universally over all allocation policies. For example, a policy might lower the ranking of a patient with a favourable blood group in anticipation of future supply of compatible livers. In this thesis the assumption is, however, justified by the bylaws of the UNOS-OPTN, as currently these do not take blood group, or future supply, into consideration for determining the ranking of patients on the waiting list. Another limitation is that our analysis of the heterogeneous treatment effect involves parametric assumptions. Although we expect changes in the treatment effect to be smooth it is not necessarily linear. Alternatives such as performing the analysis in a subset, kernel smoothing or k-nearest neighbors are more flexible, however, for this research we highly valued the ability of the interaction terms to test if the treatment effect changed relative to the average treatment effect. Although it was out of scope for this thesis, for future research we advice to extend our work by investigating the effect of a longer waiting time on dropout and formulate a unified framework in which these two outcomes could be traded off. In that framework, the current research estimated the causal impact of waiting on post-operative survival which is robustly grounded in the statistical and causal inference theory.

7 Acknowledgements

I would like to thank Dr. Naghi for supervising me with this thesis. I really appreciated your patience, excitement, and feedback. Furthermore, I would like to acknowledge Prof. IJzermans for giving me the opportunity to make my thesis at the Transplant institute of the Erasmus Medical Center and combine my econometrics master with my medical study and research. Thank you for all your, expertise, and the many enjoyable conversations we had. For me, this project, has tought me so much about causal inference. Without knowing it in advance, I think this was actually what I looked for when starting econometrics. The field is beautifully simple and complex, concrete and abstract, theoretical and applied, all at the same time. Throughout this project, I have developed a deep appreciation for the many scholars which papers I have enjoyed reading which were fundamental to make this thesis work.

I would also like to take this opportunity to thank all the teachers that have taught me over the past years. I would like to specifically thank Dr. Oldenkamp, Prof. Fok, Prof. Paap, Dr. Heij, Dr. van den heuvel, Dr. Dollevoet, Dr. Zhelonkin, and Dr. Brinkhuis for revealing for me completely hidden worlds. Most importantly, your more subtle lessons in self-confidence, humbleness, and on how to express thoughts with clarity, will always be part of me.

Without my close friends Joost, Jaron, Patrick, Lisa, Anna, Emma, and Celine, this period would not have been the same. Laughing, speaking, learning and being together was great. Thijs, thank you for your boldness, enthusiasm, and kindness. Without you, I would never have considered applying for econometrics; one of the best decisions I took in life.

Most of all, I want to thank my family. Mees and Rose you have always been there for me. Mum and dad, this was only possible because of your dedication, attention, love, and care. You helped me growing, not only in the past months or years, but for as long as I can remember, and I'm really proud of that. Ele, you know me best, and you also know that this journey has been very fun, and interesting, but also intimidating and full of self-doubt. Thank you for all your help and sharing life with me every single day.

8 List of abbreviations

Table 9: List of abbreviations

Abbreviation	Description
LT	Liver transplantation
HCC	Hepatocellular carcinoma
TT	Test-of-time
MELD score	Model of endstage liver disease score
IV	instrumental variables
ATE	average causal treatment effect
OPTN	Organ procurement and transplantation network
UNOS	United network for organ sharing
BMI	Body mass index
ALBI score	Albumine Bilirubin score
LRT	Locoregional therapy
AFP	Alpha fetoprotein
IPW	Inverse probability weighting
Educ	Education
Ethn	Ethnicity
Enceph	Encephalopathy
Tum num	Tumor number
Tum size	Tumor size
Tum size tot	Sum of tumor sizes
LF	Liver Function
2SLS	Two stage least squares
KS	Kolmogorov-Smirnov
SMD	Standardized mean difference
n	Number of patients
Med	Median
CI	Confidence Interval
OS	Overall survival

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726.

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925.

Aalen, O. O. (1993). Further results on the non-parametric linear regression model in survival analysis. *Statistics in medicine*, 12(17):1569–1588.

Andersen, P. K., Borgan, Ø., Hjort, N. L., Arjas, E., Stene, J., and Aalen, O. (1985). Counting process models for life history data: A review [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 97–158.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pages 313–336.

Angrist, J. D., Graddy, K., and Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Angrist, J. D. and Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Angrist, J. D. and Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly journal of economics*, 114(2):533–575.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica: Journal of the Econometric Society*, pages 77–83.

Clogg, C. C., Petkova, E., and Shihadeh, E. S. (1992). Statistical methods for analyzing collapsibility in regression models. *Journal of Educational Statistics*, 17(1):51–74.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. (2020). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv preprint arXiv:2001.09887*.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853.

Everhart, J. E., Lombardero, M., Detre, K. M., Zetterman, R. K., Wiesner, R. H., Lake, J. R., and Hoofnagle, J. H. (1997). Increased waiting time for liver transplantation results in higher mortality1. *Transplantation*, 64(9):1300–1306.

Freeman Jr, R. B. and Edwards, E. B. (2000). Liver transplant waiting time does not correlate with waiting list mortality: implications for liver allocation policy. *Liver Transplantation*, 6(5):543–552.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention science*, 8(3):206–213.

Greenland, S. (1996). Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*, pages 498–501.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46.

Haycock, P. C., Burgess, S., Wade, K. H., Bowden, J., Relton, C., and Davey Smith, G. (2016). Best (but oft-forgotten) practices: the design, analysis, and interpretation of mendelian randomization studies. *The American journal of clinical nutrition*, 103(4):965–978.

Heckman, J. J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics*, 115(1):45–97, IV–p58.

Howard, D. (2000). The impact of waiting time on liver transplant outcomes. *Health Services Research*, 35(5 Pt 2):1117.

HRSA/OPTN (2012). Policy 3.6 organ distribution:allocation of livers. [Online; accessed 2-July-2021].

HRSA/OPTN (2018). Hcc approval criteria policy notice. [Online; accessed 14-Okt-2021].

Lergenmuller, S. (2017). Two-stage predictor substitution for time-to-event data. Master's thesis.

Li, J., Fine, J., and Brookhart, A. (2015). Instrumental variable additive hazards models. *Biometrics*, 71(1):122–130.

Li, Q., Wu, T., Ma, X.-A., Jing, L., Han, L.-L., and Guo, H. (2018). Prognostic role of abo blood group in patients with unresectable hepatocellular carcinoma after transarterial chemoembolization. *Therapeutics and clinical risk management*, 14:991.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71.

MacKenzie, T. A., Tosteson, T. D., Morden, N. E., Stukel, T. A., and O'Malley, A. J. (2014). Using instrumental variables to estimate a cox's proportional hazards regression subject to additive confounding. *Health Services and Outcomes Research Methodology*, 14(1):54–68.

Martinussen, T. and Vansteelandt, S. (2013). On collapsibility and confounding bias in cox and aalen regression models. *Lifetime data analysis*, 19(3):279–296.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, 81(3):501–514.

Nagai, S., Kitajima, T., Yeddula, S., Salgia, R., Schilke, R., Abouljoud, M. S., and Moonka, D. (2020). Effect of mandatory 6-month waiting period on waitlist and transplant outcomes in patients with hepatocellular carcinoma. *Hepatology*, 72(6):2051–2062.

Oral, A. and Sahin, T. (2019). Prognostic role of abo blood group and rhesus factor in cirrhotic patients with hepatocellular carcinoma. *Scientific reports*, 9(1):1–6.

Pearl, J. (1993). [bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.

Pearl, J. (2009). *Causality*. Cambridge university press.

Robins, J. M. and Tsiatis, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in statistics-Theory and Methods*, 20(8):2609–2631.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Salvalaggio, P., Felga, G., Axelrod, D., Della Guardia, B., Almeida, M., and Rezende, M. (2015). List and liver transplant survival according to waiting time in patients with hepatocellular carcinoma. *American Journal of Transplantation*, 15(3):668–677.

Schlansky, B., Chen, Y., Scott, D. L., Austin, D., and Naugler, W. E. (2014). Waiting time predicts survival after liver transplantation for hepatocellular carcinoma: A cohort study using the u nited n etwork for o rgan s haring registry. *Liver Transplantation*, 20(9):1045–1056.

Sjölander, A., Dahlqwist, E., and Zetterqvist, J. (2016). A note on the noncollapsibility of rate differences and rate ratios. *Epidemiology*, 27(3):356–359.

Staiger, D. O. and Stock, J. H. (1994). Instrumental variables regression with weak instruments.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73(2):363–369.

Tchetgen, E. J. T., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)*, 26(3):402.

Theil, H. (1953). Repeated least squares applied to complete equation systems. *The Hague: central planning bureau*.

Van Buuren, S. (2018). *Flexible imputation of missing data.* CRC press.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The annals of mathematical statistics*, 11(3):284–300.

Wright, P. G. (1928). *Tariff on animal and vegetable oils.* Macmillan Company, New York.

Wu, T., Ma, X.-A., Wang, G.-Q., Li, Q., Li, M.-J., Guo, J.-Y., Liang, X., Ruan, Z.-P., Tian, T., Nan, K.-J., et al. (2017). Abo blood type correlates with survival in hepatocellular carcinoma following hepatectomy. *Scientific reports*, 7(1):1–9.

Zhong, M. and Hess, K. R. (2009). Mean survival time from right censored data.

9 Appendix A

Supplementary file: Appendix A - Preprocessing.xlsx

10 Appendix B

Convergence of multiple imputations of the variables: functional status, ALBI score, MELD score, Tumor number, Tumor size and log10(AFP).

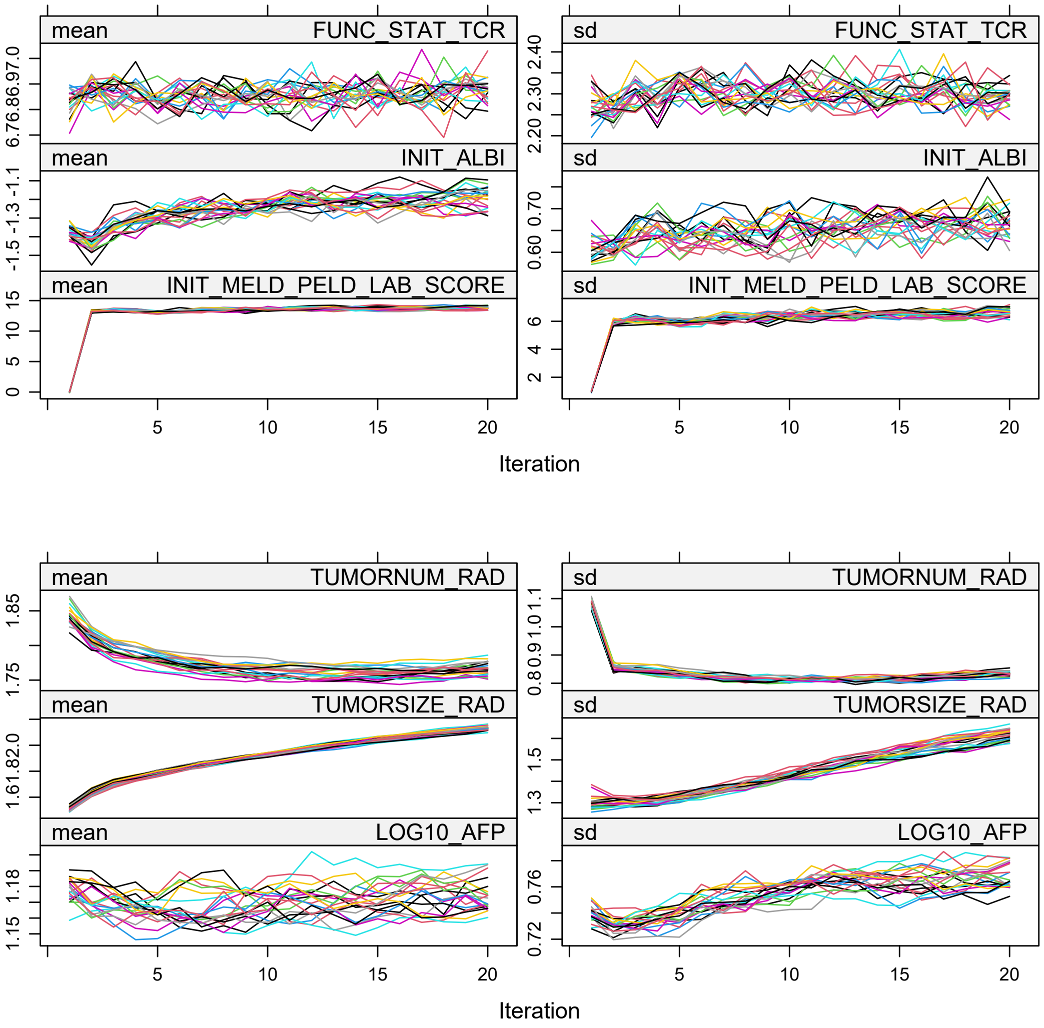

Figure 7: Convergence of imputations

11 Appendix C

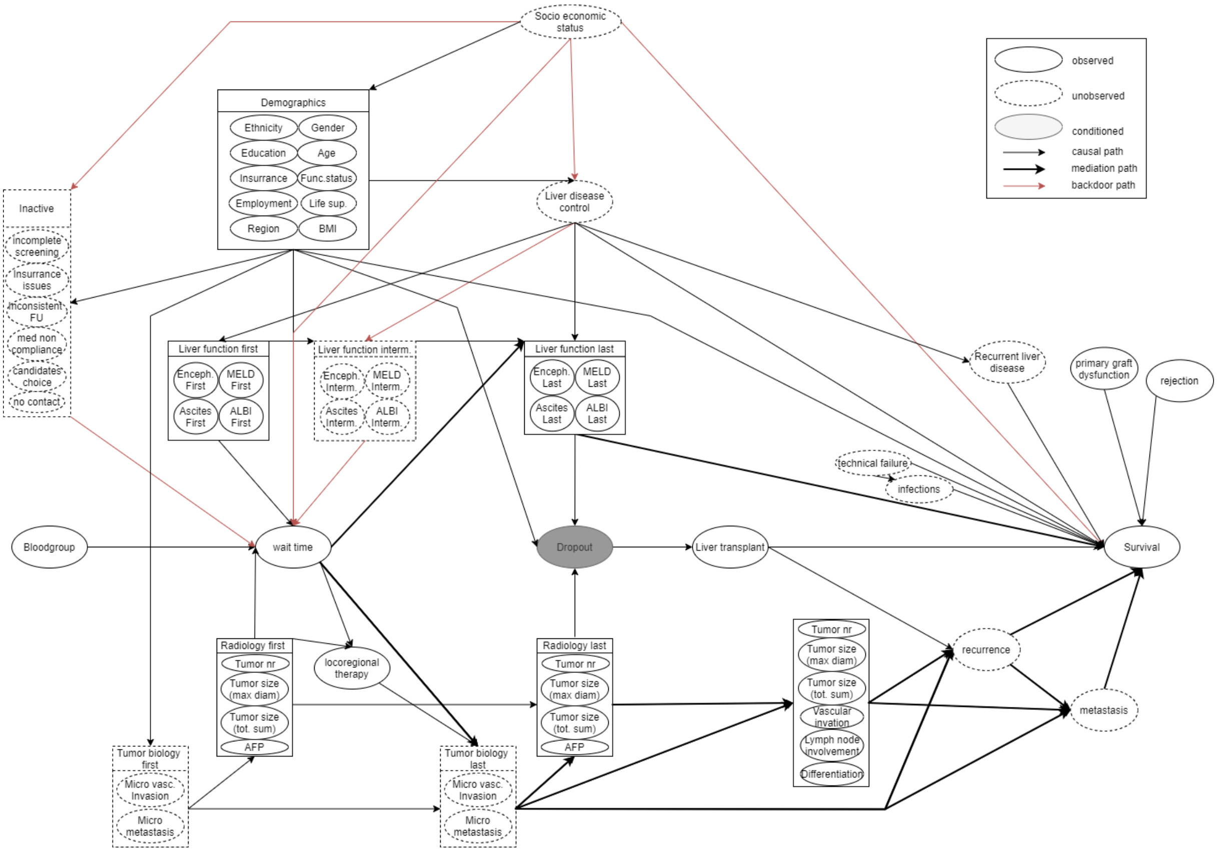

Figure 8: Detailed causal graph

12 Appendix D

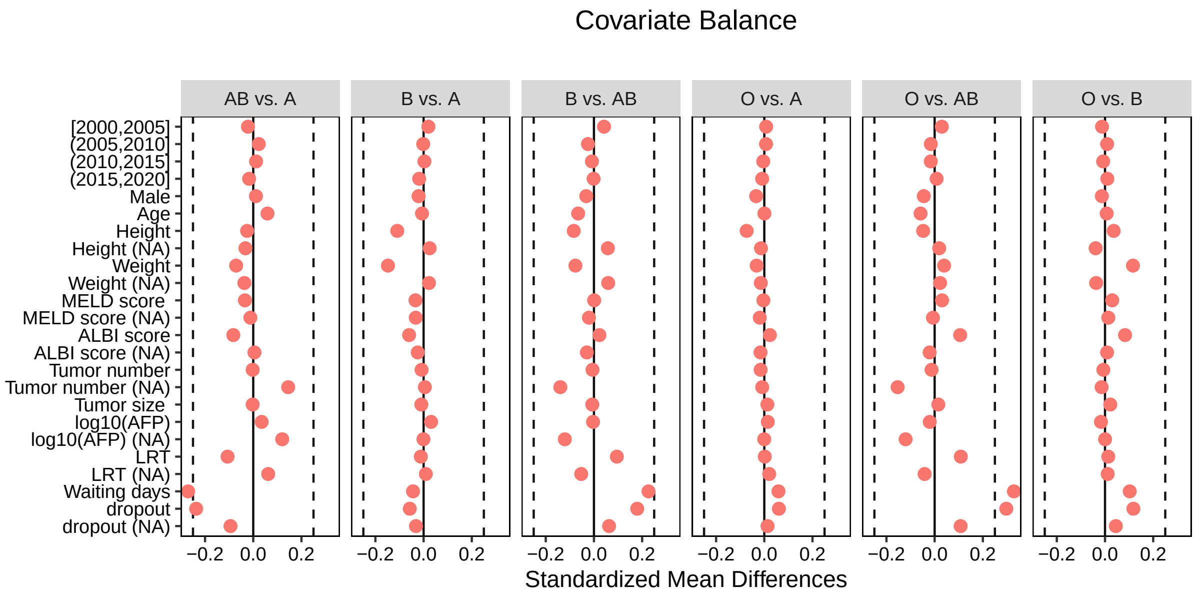

Figure 9: Covariate balance