Retrospective analysis of PFIC data

Statistical modelling of disease trajectories and survival towards a better understanding of PFIC disease

by

P. Huisman

to obtain the degree of Master of Science in Applied Mathematics

at the Delft University of Technology,

to be defended publicly on Tuesday June 24, 2025 at 13:30 AM.

Student number:	4933095		
Project duration:	August 29, 2024 – June 24, 2025		
Thesis committee:	Prof. dr. ir. G. Jongbloed,	TU Delft,	supervisor
	Prof. dr. ir. B. E. Hansen,	Erasmus MC,	external supervisor
	Ir. P. Miranda Afonso,	Erasmus MC,	external supervisor
	Dr. ir. G. F. Nane,	TU Delft	

TU Delft

Erasmus MC
University Medical Center Rotterdam

Preface

This thesis marks the final step in completing my Master's in Applied Mathematics at Delft University of Technology, with a specialisation in Stochastics. Over the past nine and a half months, I have discovered that applying mathematics and statistics to socially relevant challenges is both deeply engaging and motivating for me.

I am sincerely grateful to everyone who has supported and guided me throughout this journey. First and foremost, I would like to thank my thesis supervisors, Geurt Jongbloed, Bettina Hansen, and Pedro Miranda Afonso. Thank you for the continuous motivation that has encouraged me to navigate through this research. Bettina and Pedro, thank you for generously sharing your expertise and for introducing me to the field of Biostatistics. I am especially thankful for the opportunity to present my work at the conferences in Amsterdam and Helsinki. Your support made those experiences both successful and memorable. Geurt, I am especially grateful for your support, empathy, and constructive feedback throughout the process. Your optimism helped me stay on track, and your sense of humour made every meeting both insightful and enjoyable. I would also like to thank all of you for making the collaboration between Erasmus MC and TU Delft possible. Furthermore, I would like to thank Tina Nane for being part of my thesis committee and for taking the time to review my research.

Finally, I am deeply thankful to my friends and family for their support over the past nine and a half months. Your encouragement has been crucial in maintaining perspective.

Thank you all for being part of this journey. This thesis would not have been possible without you.

P. Huisman
Rotterdam, June 2025

Abstract

Progressive familial intrahepatic cholestasis (PFIC) is a group of rare, inherited liver diseases that affect children and are characterised by impaired bile flow. Since PFIC is a paediatric ultra-rare disease, conducting randomised controlled trials is particularly challenging, making observational data essential for improving clinical management. This thesis analyses a large multinational observational retrospective data cohort with long-term follow-up of PFIC patients. The aim is to improve our understanding of PFIC and support more informed decision-making in patient care through the investigation of two key aspects of disease monitoring and progression. First, the thesis explores longitudinal trajectories of relevant biochemical parameters, serum bile acid levels and platelet counts, in patients with a specific subtype of PFIC, PFIC2, using latent class linear mixed models. This approach effectively identified distinct longitudinal patterns of serum bile acids and platelet counts in patients with PFIC2. These patterns highlight significant heterogeneity in the progression of laboratory parameters over time. Second, a comparative analysis of event-free survival is conducted between two European regional cohorts of PFIC patients, North-West Europe and South-Central Europe. Hypothesising that there are no differences in event-free survival of PFIC patients despite different care settings. This is achieved through a weighted survival analysis combining inverse probability treatment weighting with the Kaplan-Meier estimator and the Cox proportional hazards model. The results suggest there are no significant regional differences in event-free survival among PFIC2 patients between the two cohorts. Furthermore, a sensitivity analysis and permutation test have been performed, which also support this result. Together, these findings contribute to a more detailed understanding of disease progression in PFIC patients and provide practical tools and insights that can inform patient monitoring and clinical decision-making in the absence of randomised trials.

Contents

Abbreviations

Abbreviation	Definition
AIC	Akaike's Information Criterion
ALT	Alanine Transaminase
BIC	Bayesian Information Criterion
BSEP	Bile Salt Export Pump
CI	Confidence Interval
HR	Hazard Ratio
IPTW	Inverse Probability of Treatment Weighting
KM	Kaplan-Meier
LCLMM	Latent Class Linear Mixed Model
LTx	Liver Transplantation
NAPPED	NAtural course and Prognosis of PFIC and the Effect of biliary Diversion
PFIC	Progressive Familial Intrahepatic Cholestasis
PH	Proportional Hazard
PLT	Platelet Count
sBA	Serum Bile Acids
SBD	Surgical Biliary Diversion
SMD	Standardized Mean Difference
ULN	Upper Limit of Normal

1

Introduction

This thesis presents an analysis of observational data retrospectively collected from patients with the ultra-rare liver disease *progressive familial intrahepatic cholestasis (PFIC)*. PFIC primarily affects children and, like many rare diseases, poses major challenges for conducting randomised controlled trials due to ethical concerns, high costs, limited patient populations, and time constraints. In the absence of randomised trials, observational data serve as a valuable resource for gaining clinical insights and informing treatment strategies. The observational data used in this thesis are from a global cohort of 76 sites worldwide with long-term follow-up of patients with PFIC.

This thesis investigates two key aspects of understanding and monitoring PFIC. First, it models the longitudinal trajectories of biochemical parameters to explore disease progression in PFIC patients. Second, it performs a comparison of event-free survival, defined as the time until a patient experiences their first clinical event (e.g. death), between two regional cohorts of PFIC patients within Europe. Together, these analyses aim to improve our understanding of PFIC and to support more informed decision-making in patient care.
The disease PFIC is first introduced in Section 1.1, after which the motivation and research objectives of this thesis are outlined in Section 1.2.

1.1. Progressive familial intrahepatic cholestasis

PFIC is a heterogeneous group of inherited liver diseases that primarily affect children. It causes a buildup of bile in the liver, which is typically observed early in newborns or within the first year of life. Bile is toxic, so this accumulation can lead to cholestasis, severe liver damage, and, in many cases, liver failure, which typically occurs between infancy and adolescence. PFIC is classified into different subtypes, each caused by genetic mutations that affect how liver cells transport bile, which is important for its normal production and flow [12]. The most common subtypes are PFIC1 and PFIC2, which together account for the majority of PFIC cases; therefore, this thesis focuses on these two types. Although the exact incidence of PFIC remains uncertain, it is recognised as an ultra-rare disease, with estimates ranging from 1 in 50,000 to 1 in 100,000 live births [12]. In the Netherlands, this translates to 2 to 3 cases per year, based on figures from CBS [9].

Cholestasis, a key clinical feature of PFIC, results from impaired bile flow. It is characterised by jaundice and pruritus in early childhood. Among PFIC symptoms, pruritus is the most debilitating, particularly in patients with PFIC1 and PFIC2 [40]. For patients and their families, it poses a considerable burden [29]. Severe pruritus can lead to skin damage (often with bleeding), sleep disturbances, growth problems, irritability, difficulty concentrating, and poor school performance. Although the exact cause of cholestatic pruritus remains unclear, it appears to be related to increased concentrations of the biochemical parameter serum bile acids (sBA) [30]. Other biochemical parameters may also reflect the severity of the disease, such as platelet counts, alanine transaminase (ALT) and total bilirubin.

PFIC progresses rapidly to fibrosis and end-stage liver disease. Without treatment, end-stage liver disease is ultimately fatal [29]. Management of PFIC involves both medical and surgical approaches,

with the primary goal of relieving pruritus. Treatment includes the use of medications as initial therapy to relieve the pruritus [40]. When these are insufficient, surgical options are considered, most commonly surgical biliary diversion (SBD). SBD aims to reduce the recycling of bile acids between the liver and intestines, thereby lowering toxic bile salt accumulation in the body [40]. SBD is a major surgical procedure, and the child will be living with a stoma for ongoing care afterwards. Studies have shown that sBA concentration is reduced in patients who respond well to SBD [39]. For 23–75% of patients undergoing SBD, no further surgical intervention is required [29]. Of those who do not benefit, many ultimately require a liver transplantation. Liver transplantation is indicated for patients who do not respond to medical or surgical treatment, have developed end-stage liver disease, or suffer from severely impaired quality of life due to persistent uncontrolled pruritus [29]. It has been shown to resolve cholestasis and improve symptoms in 75–100% of patients, regardless of PFIC subtype, within a short-term follow-up period of 3 to 5 years [40].

1.2. Motivation and Research Objectives

In this thesis, we hypothesise that identifying patient subgroups based on the longitudinal trajectories of biochemical parameters may contribute to a better understanding of PFIC and support more personalised treatment approaches. Specifically, recognising which patients are at a higher risk could guide the development of targeted therapeutic interventions, as well as enable earlier and more effective treatment. To identify such subgroups, we focus on two potentially relevant biochemical markers: sBA levels and platelet counts, both of which are known to be associated with disease severity. This analysis is conducted using a latent class linear mixed model. This leads us to the first objective of this thesis:

1. Determine and identify similarities of trajectories of the relevant biochemical parameters, sBA levels and platelet counts in patients with PFIC2.

The clinical relevance of this analysis is substantial, as understanding these trajectories may inform both prognosis and treatment strategies. The impact and significance of this research on the medical community are highlighted by the acceptance of the results for presentation at two clinical conferences: the European Association for the Study of the Liver (EASL) Congress 2025 in Amsterdam (May 2025) and the 57th Annual Meeting of the European Society for Paediatric Gastroenterology, Hepatology, and Nutrition (ESPGHAN) in Helsinki (May 2025). The poster presented at EASL is given in the Appendix in Figure C.1.

Furthermore, although the dataset includes patients around the world, a substantial proportion of PFIC originates from Europe. There may be regional differences in the care settings, how PFIC is diagnosed, evaluated, or treated across hospitals and countries. Investigating whether these differences affect patient outcomes could provide valuable insights. This thesis hypothesises that despite different care settings, there are no differences in event-free survival in PFIC patients, where event-free survival is defined as the time until a patient experiences their first clinical event, which in this context includes liver transplantation, SBD, or death. Therefore, this thesis also explores the potential effect of geographic region on event-free survival. The analysis compares two regional cohorts: North-West Europe and South-Central Europe. This leads to the second objective of this thesis:

2. To perform a comparison of event-free survival between two regional cohorts of PFIC patients within Europe.

To be able to perform a fair comparison between two cohorts in an observational study, some control for confounding will be performed.

1.3. Thesis structure

The remainder of this thesis is structured as follows. Chapter 2 introduces the observational dataset used in this study, including a detailed description of the variables and relevant clinical variables. Chapter 3 presents the methodology for conducting the comparison of event-free survival. This includes an overview of survival analysis techniques and the approach used to adjust for confounding, the inverse probability of treatment weighting. The chapter concludes by integrating these components into the final analytical framework. Chapter 4 addresses the first objective by introducing the latent class lin-

ear mixed model, explaining its application, and presenting the resulting patient subgroups based on longitudinal biochemical data. Chapter 5 then focuses on the second objective, presenting the results of the event-free survival comparison between regional cohorts, as outlined in Chapter 3. Chapter 6 concludes with a discussion of the findings, a reflection on the study's limitations, and suggestions for future research.

2

Data

As mentioned in Section 1.1, PFIC is recognised as an ultra-rare disease, meaning it affects only a relatively small number of individuals. As a consequence, limited information is available on the natural history of PFIC. To improve the understanding of its natural history, phenotype variability, and the association between treatments and long-term outcomes, an international multicenter initiative was launched in 2017: the NAtural course and prognosis of PFIC and the effect of biliary diversion (NAPPED) initiative [46].

Data were retrospectively collected from 76 centres worldwide. The consortium was initiated by the Department of Paediatrics at the Beatrix Children's Hospital, University Medical Centre Groningen (UMCG). Researchers at each participating centre gathered clinical, demographic, and outcome data by identifying all consecutive patients under 18 years of age who had received pediatric care since 1977 [47]. As of the data export date for this study, September 2024, the NAPPED dataset included 1,010 patients with a total of 7,201 measurements.

The data are stored in a longitudinal format. The general structure of this dataset is given in Table 2.1. The variable $x_{i,j,k}$ is the k-th clinical, demographic or outcome value for the j-th visit of the i-th patient, with $1 \leq i \leq I$, $1 \leq j \leq n_i$ and $1 \leq k \leq q$. There are $q = 367$ variables collected per patient visit in the NAPPED database.

Table 2.1: Structure of the longitudinal data

Patient ID	Visit	Variables		
1	1	$x_{1,1,1}$	$\dots$	$x_{1,1,q}$
1	2	$x_{1,2,1}$	$\dots$	$x_{1,2,q}$
.	.	.	.	.
1	n_1	$x_{1,n_1,1}$	$\dots$	$x_{1,n_{1,q}}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
I	1	$x_{I,1,1}$	$\dots$	$x_{I,1,q}$
I	2	$x_{I,2,1}$	$\dots$	$x_{I,2,q}$
.	.	.	.	.
I	n_I	$x_{I,n_I,1}$	$\dots$	$x_{I,n_I,q}$

Demographic variables in this study include sex, region, date of birth, etc. Each visit gives the date of that specific visit, using this information, the age can be calculated. Clinical variables are, among others, biochemical parameters for liver health. Every visit captured in the dataset gives the laboratory measurements of some clinical variables; when a measurement of a certain factor is not taken at a specific visit, the value is given as not available (NA). The data also gives the 'upper limit of normal (ULN)' of the biochemical parameters used in this thesis. The ULN is a crucial threshold used to define the highest value of a measurement that is considered within the normal range for a healthy individual [25]. The ULN is different per laboratory, since laboratories may use different analytical methods or

instruments to measure the same biomarker [14]. This value is frequently used in clinical trial protocols to define inclusion/exclusion criteria.
The outcome data are the information about the events of the patients during the time they have been followed. These are factors which indicate, for example, whether the patient underwent a SBD, a liver transplantation or if they died, including the dates of these events.

This thesis primarily focuses on patients with PFIC2. For the analysis of longitudinal patterns in two biochemical parameters, only the data from PFIC2 patients are used. For the comparison of event-free survival between two regional groups, the data from both PFIC1 and PFIC2 patients are analysed separately, but only for the patients in Europe. Table 2.2 presents selected characteristics of the patients in the regional groups to provide an overview of the study population. In both analyses, the data are filtered based on specific selection criteria, detailed in the corresponding chapters. These chapters also provide more detailed information about the characteristics of the selected data.

Table 2.2: Characteristics of the study population.

		North-West Europe (EU N/W)		South-Central Europe (EU S/C)	
		PFIC1 (n = 59)	PFIC2 (n = 228)	PFIC1 (n = 35)	PFIC2 (n = 136)
Sex:	Male: n (%)	38 (64.4)	97 (42.5)	25 (71.4)	79 (58.1)
	Female: n (%)	20 (33.9)	125 (54.8)	10 (28.6)	55 (40.4)
	Unknown: n (%)	1 (1.69)	6 (2.63)	0 (0)	2 (1.47)
Age at diagnosis (first visit): median (IQR)		0.55 (0.33–1.43)	0.83 (0.24–2.26)	0.58 (0.24–2.26)	0.83 (0.34–1.92)
SBD: n (%)		24 (40.7)	59 (25.9)	17 (48.6)	33 (24.3)
Liver transplantation: n (%)		25 (42.4)	113 (49.6)	21 (60.0)	62 (45.6)
Death: n (%)		7 (11.9)	18 (7.89)	2 (5.71)	5 (3.68)
Years of follow-up: median (IQR)		4.75 (1.42–11.1)	6.25 (2.19–12.4)	4.88 (1.84–8.05)	5.42 (1.94–13.5)
Number of visits per patient: median (IQR)		7 (3–13)	9 (5–16)	7 (6–12)	8 (4–16)

3

Methodology for the comparison of event-free survival between two cohorts

One of the objectives of this thesis is to perform a comparison of the survival time distribution until the first event of liver transplantation, SBD or death of the PFIC1 and PFIC2 patients in the two divided regions in Europe, to study the regional effect on the survival rate of PFIC. Before proceeding with the results of this comparison, we must first understand the methodology for performing this comparison. This chapter focuses on explaining this methodology. We begin by introducing the survival analysis research area. This is needed to calculate the survival time distributions and eventually perform the comparison between the two groups.

However, we are focusing on data from an observational study. A fair comparison between two groups in an observational study is not as straightforward due to measured and unmeasured differences in characteristics between groups, because of the lack of randomisation [10]. To control for confounding in observational studies, the inverse probability treatment weighting (IPTW) method is used. This statistical method assigns a specific weight to each individual in the dataset, based on how representative they are of the overall population. Individuals from underrepresented subgroups are given higher weights, thereby increasing their contribution to the analysis. By doing so, the method creates a pseudo-population in which the measured confounders are more equally distributed among groups [10]. This statistical method is explained in Section 3.2.

To eventually perform the comparison of survival time distributions, the two statistical methods, survival analysis and IPTW, are combined in Sections 3.3 and 3.4.

3.1. Survival Analysis

Survival analysis, generally called analysis of time-to-event data, is a collection of statistical techniques for data analysis where the outcome variable of interest is **time until an event occurs**. One of the goals in survival analysis is to predict the expected duration until an event occurs. In a medical setting, the event is often clinical and could be, for example, death, disease occurrence, recovery from the disease, surgery, or any experience that happens to an individual [23]. The time until an event occurs is often referred to as the survival time, since a common event in these types of studies is the death of an individual.

In survival analysis, two functions are important and often used. These are the survival function and the hazard function. The terminology is inspired by the situation where the clinical event is death. The survival function gives the probability that an individual is still alive after some specified time.

Definition 3.1.1 (Survival function [22]) *Let X be a nonnegative random variable denoting an event*

(survival) time. The survival function of X is defined as

$$S(t) = P(X > t) = 1 - F(t), \text{for } t \geq 0 \tag{3.1}$$

where F is the distribution function of X. Note that $S(t)$ is non-increasing function, $0 \leq S(\infty) \leq S(0) \leq 1$.

The hazard function gives "the instantaneous potential for the event to occur, given that the individual has survived up to time t" [23]. In other words, the hazard function reflects the danger (hazard) of getting killed at a specific time, having survived till (just before) this time [22]. Thus, the hazard function is also called the failure rate.

Definition 3.1.2 (Hazard function [22]) *The hazard function of a discrete random variable X is defined as*

$$h(t) = P(X = t | X \geq t) = \frac{P(X = t)}{1 - F(t-)} \tag{3.2}$$

$F(t-)$ denotes the left hand limit of F at t.
The hazard function of a continuous random variable X with density f is defined as

$$h(t) = -\frac{d}{dt} \log S(t) = \frac{f(t)}{S(t)} \tag{3.3}$$

The hazard function is the framework through which mathematical modelling of survival data is implemented [23]. Based on equation 3.3, the survival function can be expressed in terms of the cumulative hazard function H, where $H(t) = \int_{[0,t]} h(s)ds$. Indeed, $S(t) = \exp(-H(t))$.

The basic goals of survival analysis are to estimate and interpret survival - and hazard functions from survival data, to compare survival and hazard functions, and to assess the relationship of explanatory variables to the distribution of the survival time [23]. To reach these goals, especially the first and second goals, the survival function needs to be estimated. This is usually carried out using the Kaplan-Meier method, which will be explained in Section 3.1.2. The third goal also requires some mathematical modelling. To assess the relationship between the survival time and one or more explanatory variables, the Cox proportional hazard regression model is most commonly used, explained in Section 3.1.4. The explanatory variables may be variables that are specified by the experimental conditions, such as receiving a treatment, or they may be observational variables, such as the region where the patient lives. The latter is the explanatory variable that is focused on in this thesis.

3.1.1. Censoring

A crucial concept in survival analysis is censoring. Censoring occurs when the precise survival time is not available for certain individuals. In the medical field, for example, in a clinical study, an individual often gets lost to follow-up. This means that they take part in the study but suddenly disappear from it [22]. If one wants to analyse the time until a certain event, but the individual got lost to follow-up or did not experience the event when the study ended, this leads to right-censored data. So, right-censored means that the true survival time is equal to or greater than the actual observed follow-up time. [23]

The right-censoring problem is modelled as follows [22]. Suppose $T_1, T_2, \ldots, T_n$ are i.i.d. survival times with $C_1, C_2, \ldots, C_n$ the censoring times, i.i.d. of T_i and finally $Y_1, Y_2, \ldots, Y_n$ the actual observed follow-up times. Suppose that the observations are denoted as (Y_i, δ_i) for $i = 1, 2, \ldots, n$, where

$$Y_i = \min\{T_i, C_i\}$$

and

$$\delta_i = I(T_i < C_i).$$

Here, δ_i indicates whether the actual observed follow-up time is due to the event or censoring.

3.1.2. Kaplan-Meier estimator

Estimating the survival function can be carried out using the Kaplan-Meier (KM) method. The Kaplan-Meier estimator is the nonparametric maximum likelihood estimator in the right-censoring model [22].

Suppose $\{t_1, t_2, \ldots, t_n\}$ are the survival times, with $m \leq n$ distinct values which are denoted by $u_1 < u_2 < \cdots < u_m$. The derived maximum likelihood estimator of the survival function, called the Kaplan-Meier estimator, is given in equation 3.4.

Definition 3.1.3 (Kaplan-Meier estimator [22])

$$\hat{S}(u_i) = \prod_{j=1}^{i} \left(1 - \frac{d_j}{r_j}\right), \tag{3.4}$$

where d_j is the number of events at time u_j, and r_j is the number of individuals still at risk at time u_j, which means that the individuals are still alive and are not censored at time u_j.

The data used in calculating the Kaplan-Meier estimates need to consist of

- the observed follow-up time per patient,
- the status of the patient at the end of their follow-up, whether an event has occurred, or is censored,
- and the study group the patient belongs to [35].

The data is originally in a longitudinal format. This means that each patient can have multiple visits with measured information. The format is given in Table 2.1. The relevant data need to be extracted since we only need the above information for each patient. The follow-up times are calculated as follows

- The patient experienced an event: the follow-up time is the time between the start of the follow-up and the date of the first event of interest, in years (survival time).
- The patient is censored from the follow-up: the follow-up time is the time from the start of the follow-up until the date of censoring, in years (censoring time).

The mathematical description of the newly created dataset used in this thesis is given in Section 5.2. The start of the follow-up, called the index time or baseline, has to be chosen. When we would have a study where everyone starts the study at the same time, the index time would be the date of the start of the study. But this is not the case in our data. What the opportunities are to choose this index time and how to do this is explained later in Section 5.1.2.

The KM survival probabilities are plotted in a Kaplan-Meier curve. An example of a Kaplan-Meier curve is shown in Figure 3.1. Here, the data of the European patients of the NAPPED dataset is used. The empirical plot is plotted as a step function, which steps down to the survival probabilities as we move from one ordered survival time to another [23]. Censored individuals are indicated on the Kaplan-Meier curve as tick marks, these are the '+' signs on the plot in Figure 3.1.

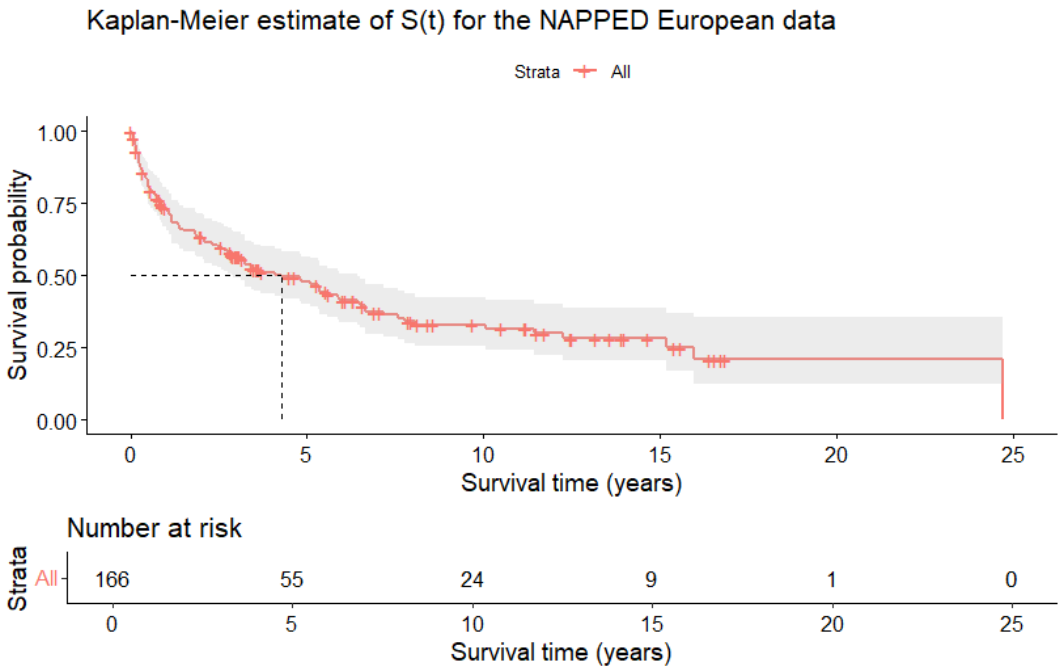


Figure 3.1: Kaplan-Meier curve for the European patients in the dataset

The estimated survival function in the Kaplan-Meier curve shows that, for example, the probability of surviving 10 years is 34%. The estimated survival function can be used to extract estimates of specific percentiles of interest, such as the median survival time. In the estimated survival function of Figure 3.1, the median survival time is 4.8 years, indicated with the dotted line.
The table below the Kaplan-Meier curve in Figure 3.1 shows the number of patients that were still at risk at a certain time. This time is given as the survival time, so it is counted from the start of the follow-up per patient, the index time. These patients, who are still at risk, are those who haven't yet experienced the event of interest and are not censored. Thus, censoring removes the patient from the individuals at risk, r_i in equation 3.4, which means that censoring affects the survival rates [35].

In the absence of censoring, the Kaplan-Meier estimator is an estimate of the survival function, it corresponds to one minus the empirical distribution function. However, once censoring occurs, the estimation becomes less direct, as it is no longer possible to determine whether censored patients would have experienced the event at a later time. Consequently, an increasing number of censored observations can reduce the reliability and precision of the survival curve [35]. Nevertheless, it remains essential to include censored patients in the analysis, as their survival times up to the point of censoring provide valuable information and contribute meaningfully to the overall estimation of the survival function.

The shaded area in the Kaplan-Meier curve is the 95% pointwise confidence interval for S(t). The Greenwood formula is used to estimate the standard error of the Kaplan-Meier survival estimate [23]:

$$\mathrm{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{r_i\,(r_i - d_i)} \tag{3.5}$$

Then, a log-transformation is applied to compute the 95% pointwise confidence interval [42]:

$$\log \hat{S}(t) \pm 1.96\sqrt{\mathrm{Var}[\log \hat{S}(t)]}.$$

3.1.3. Comparing survival functions

In clinical trials, the goal is often to compare different groups of patients in terms of survival. For example, treated versus untreated (placebo) patients, female versus male, etc. To be able to perform these comparisons, separate survival curves are estimated for the 2 groups. An example of a Kaplan-Meier plot for two groups is given in Figure 3.2.

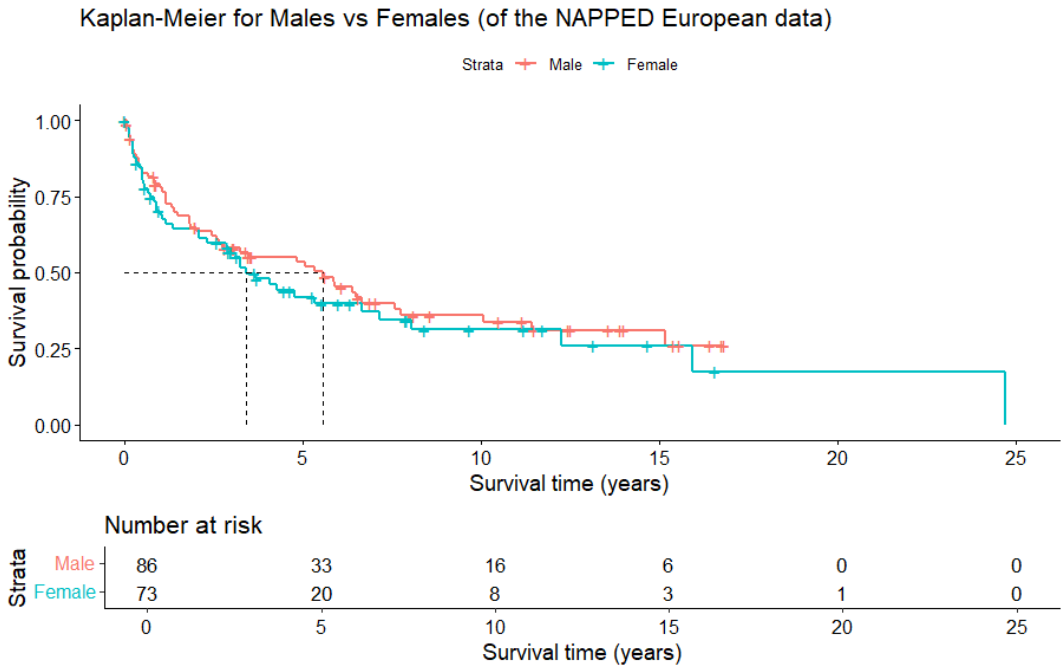


Figure 3.2: Two Kaplan-Meier curves for the comparison of males versus females in the European patients of the dataset

To compare the different groups, the following hypotheses are tested. The hypothesis for this problem is as follows [36]. Let $S_A(t)$ and $S_B(t)$ denote the survival functions for the two compared groups A and B, respectively.

H_0 : The distribution of survival times is the same for the two compared groups:
$S_A(t) = S_B(t)$ for all $t \geq 0$

H_1 : The distribution of survival times is not the same for the two compared groups:
$S_A(t) \neq S_B(t)$ for at least one $t \geq 0$.

To assess the validity of the null hypothesis, the survival curves are compared over the whole follow-up period, and the difference needs to be quantified [35]. Statistical tests, which reject or accept a null hypothesis, are used to test the difference in the distribution of survival times. The most famous statistical test to test these kinds of hypotheses is the Log-Rank test, which is also used in this thesis.

The Log-Rank test is a nonparametric test, which means that no assumption is made about the distribution of the survival times of the two compared groups [23]. The idea behind this test is to construct 2×2 contingency tables for each unique survival time and compare observed with expected numbers of events.

	Group 1	Group 2	Total
Event	d_{1i}	d_{2i}	d_i
No Event	$r_{1i} - d_{1i}$	$r_{2i} - d_{2i}$	$r_i - d_i$
At risk	r_{1i}	r_{2i}	r_i

Denote, for each ordered survival time $u_{(i)}$, d_{ji} as the number of individuals who had the event at that time in group j, and r_{ji} as the number of individuals at risk at that time in group j [23]. Let d_i and r_i be the total number of individuals who had the event and were at risk, respectively.

Under the null hypothesis, the survival curves of the two groups are the same, the expected number of individuals who had the event at time u_j can be estimated as:

$$\hat{E}_{ji} = \frac{d_i r_{ji}}{r_i} \tag{3.6}$$

With the variance of $\hat{E}_{ji}$ estimated as [23] [36]:

$$\text{Var}(\hat{E}_{ji}) = \frac{r_{1i} r_{2i} d_i (r_i - d_i)}{r_i^2 (r_i - 1)} \tag{3.7}$$

This 2×2 contingency table is constructed for every observed survival time $u_{(1)}, \ldots, u_{(m)}$. Then, the log-rank statistic is computed as described in:

$$X^2 = \frac{\left(\sum_{i=1}^{m} d_{1i} - \hat{E}_{1i} \right)^2}{\sum_{i=1}^{m} \text{Var}(\hat{E_{1i}})} \tag{3.8}$$

Under the null hypothesis, the log-rank statistic X^2 is approximately standard normally distributed [23].

The log-rank test comparing survival distributions by sex in the NAPPED European dataset (Figure 3.2) results in a test statistic of $X^2 = 0.499$ with a p-value of $p = 0.48$. As the p-value exceeds the conventional significance level of 0.05, we do not reject the null hypothesis. This suggests that there is no difference in survival distribution between the two groups, the male and the female.

3.1.4. Cox Proportional Hazards Model

If one would like to compare two groups, such that one can investigate what the effect of a certain covariate is, but wants to control for other variables, the Cox Proportional Hazards Model is commonly used. This is the most well-known semiparametric regression model for survival data. The purpose of this model is to simultaneously evaluate the effect of several covariates on survival [36].

In the Cox proportional hazard model, an individual i has a hazard rate at time t of

$$h_i(t, \boldsymbol{X}) = h_0(t) \exp\left(\sum_{l=1}^{p} \beta_l X_{il} \right), \tag{3.9}$$

where

- $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$ denotes the vector of p covariates for an individual,
- $h_0(t)$ denotes the baseline hazard,
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ denotes the model parameters.

The baseline hazard function h_0 is the hazard of an event when all covariates $\boldsymbol{X}$ or all β_l's in $\boldsymbol{\beta}$ are equal to zero. So, the baseline hazard function represents the instantaneous risk of experiencing the event at t in the absence of any covariate effects. When covariates are introduced, those with a beneficial effect will reduce the overall hazard relative to $h_0(t)$, while those with a harmful effect will increase it [36].

An important property of the Cox model is that the baseline hazard $h_0(t)$ does not need to be specified if one wants to estimate $\boldsymbol{\beta}$. This is a reason why the Cox model is popular [23].

The parameters of the Cox model are estimated using the maximum partial loglikelihood, which is obtained in the following way [36].
The probability that exactly the individual i had the event at time t, conditionally on the covariates of this individual, is approximately equal to [22] [8]:

$$\frac{h_i\left(t \mid \mathbf{X}_i\right)}{\sum_{j \in R(t)} h_j\left(t \mid \mathbf{X}_j\right)} = \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta}}}{\sum_{j \in R(t)} e^{\mathbf{X}_i^T \boldsymbol{\beta}}}, i \in R(t),$$

where $R(t)$ is the set of individuals at risk at time t. The partial log-likelihood is defined as the sum of

the log probabilities of these individual contributions for events

$$\begin{aligned}\ell(\beta) &= \log L(\beta) \\ &= \log\left(\prod_{i=1}^{n}\left[\frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}}}{\sum_{j\in R(T_i)} e^{\mathbf{X}_i^T\boldsymbol{\beta}}}\right]^{\delta_i}\right) \\ &= \sum_{i=1}^{n}\delta_i\left((\mathbf{X}_i^T\boldsymbol{\beta}) - \log\left(\sum_{j\in R(T_i)} e^{\mathbf{X}_i^T\boldsymbol{\beta}}\right)\right)\end{aligned} \tag{3.10}$$

where

- δ_i is an event indicator
- $\mathbf{X}_i^T\boldsymbol{\beta} = \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$

The maximum partial likelihood estimates $\hat{\beta}$ obtained are asymptotically normally distributed [36]:

$$\hat{\beta} \sim \mathcal{N}\left(\beta_0, \{\mathcal{I}_p(\beta_0)\}^{-1}\right), \tag{3.11}$$

where

- β_0 represents the true values of the parameter vector $\boldsymbol{\beta}$
- $\{\mathcal{I}_p(\beta_0)\}$ is the information matrix derived from the partial likelihood.

In our setting, the vector of covariates added to the Cox model $\boldsymbol{X}$ consists of a single binary variable x, $\boldsymbol{X} = (x)$:

$$x = \begin{cases} 1 & \text{Region North-West Europe} \\ 0 & \text{Region South-Central Europe} \end{cases}$$

As mentioned before, the Cox model can account for confounders. The covariates for which the model needs to be adjusted are then augmented to the vector $\boldsymbol{X}$. Then equation 3.9, for $\boldsymbol{X} = (X_1, X_2)$ with X_1 the binary variable as above and X_2 a continuous adjustment variable, will change to

$$h(t) = h_0(t)e^{\beta_1 X_1 + \beta_2 X_2} \tag{3.12}$$

where $\exp(\beta_1)$ represents the comparison of the hazard of the event for individuals with $X_1 = 1$ to those with $X_1 = 0$, while holding the continuous variable X_2 constant.
In general, one-unit change in variable X_j, where $j = 1, \ldots, p$, corresponds to a β_j change of $\log\{h_i(t)/h_0(t)\}$, and thus increases $h_i(t)/h_0(t)$ by a factor of $\exp(\beta_j)$. If $\beta_j < 0$, then $\exp(\beta_j) < 1$, from which it follows that the risk will decrease [36].

Assumptions Cox model: Proportional hazards assumption

As mentioned before, the main reason to use the Cox model is that there is no assumption needed for the distribution of the baseline hazard when estimating β. The Cox Model does make an assumption, the proportional hazards (PH) assumption. The PH assumption requires that the hazard ratio of two individuals is constant over time, such that the effect of a covariate on the risk of an event is constant over time [23].

To understand the principle of the proportional hazards assumption, the formula for the hazard ratio is derived as follows [23].

$$\begin{aligned}\hat{HR} &= \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} \\ &= \frac{\hat{h_0}(t)\exp\left[\sum \hat{\beta}_i X_i^*\right]}{\hat{h_0}(t)\exp\left[\sum \hat{\beta}_i X_i\right]} \\ &= \exp\left[\sum_{i=1}^{p}\hat{\beta}_i(X_i^* - X_i)\right] = \hat{\theta}\end{aligned} \tag{3.13}$$

where $\mathbf{X}^* = (X_1^*, X_2^*, \ldots, X_p^*)$ and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ denote the set of covariates $\boldsymbol{X}$ for the two individuals, and $\hat{\theta}$ does not depend on time. This derivation shows that the final expression for the hazard ratio does not depend on time.

If the proportional hazard assumption is not satisfied, because the hazard ratio varies with time, it is inappropriate to use a Cox PH model [23]. Therefore, it is important to assess the validity of the proportional hazard assumption. The most common ways to do so are visual assessment of KM curves, log(-log) plots and testing of scaled Schoenfeld residuals [26]. A log(-log) survival curve is a transformation of an estimated survival curve that results from taking the natural log of an estimated survival probability twice [23].

3.2. Weighting methods to control for confounding

As mentioned in the Introduction, Chapter 1, randomised trials are not feasible in studies involving children with a rare disease, such as PFIC. Therefore, we rely on observational data. In randomised controlled trials, the randomisation ensures that two cohorts are comparable, both in terms of measured and unmeasured baseline characteristics. This comparability allows for a more reliable estimation of the causal effect, as any observed differences in outcomes are more likely to result from the exposure rather than from pre-existing differences.

However, in non-randomised studies, such as observational studies, there is a risk of exposure-selection bias, since individuals are not assigned randomly but instead are assigned based on their characteristics. As a result, the two groups may differ systematically in ways that are also related to the outcome [5]. This creates a major challenge for causal inference, as any observed association between exposure and outcome may be confounded by these pre-existing differences, rather than reflecting a true causal relationship.

Confounding occurs when one aims to determine the effect of an exposure on the occurrence of a disease, but then actually measures the effect of another factor, a confounding factor [21]. This can lead to an over- or underestimation of the true causal effect. The confounding factor is an external variable that influences the variables under study, leading to misleading conclusions about their relationship [32].

Due to confounding, the real effect of a variable that you want to determine is disturbed, which is why confounding needs to be controlled as much as possible. Several statistical methods have been developed for this purpose. One widely used approach in the context of causal inference is the propensity score method. In particular, inverse probability of treatment weighting (IPTW) is a technique that uses estimated propensity scores to create a so-called pseudo-population in which the distribution of baseline covariates is comparable for the exposed groups [10]. This method, used in this thesis, helps to reduce the confounding bias and supports a more valid causal interpretation of the exposure effect. The next section will introduce the IPTW method in more detail.

3.2.1. Inverse probability treatment weighting

IPTW involves two main steps. First, the propensity score - the probability that an individual is subjected to a particular exposure (e.g. receiving treatment or geographic region) given their observed characteristics - is estimated. In the second step, each individual is assigned a weight equal to the inverse of the probability of receiving the treatment they actually received [10]. This weighting process creates a pseudo-population in which the distribution of measured confounders is balanced between the exposed and unexposed groups, thereby approximating the conditions of a randomised experiment. In the created pseudo-population, the weights conceptually do represent not only the individual itself but actually w individuals, given that w represents the weight [10]. So in this pseudo-population, based on observed characteristics, some individuals are up-weighted and some down-weighted. Individuals with a lower probability of exposure are assigned larger weights, thereby increasing their influence in the weighted comparison [10].

Propensity scores

The purpose of propensity scores in observational research is to adjust for measured confounders by ensuring balance in characteristics between the groups to be compared [10].
The propensity score was first defined by Rosenbaum and Rubin [38] and has been used by many researchers to estimate the treatment effects in observational studies [49] [1]. The propensity score is defined as the conditional probability of receiving a particular exposure, given a set of observed covariates [38]. The term ”exposure” can refer to various forms of assignment, such as receiving a specific treatment, belonging to a certain sex or residing in a particular region. In this thesis, the exposure of interest is defined as belonging to a particular region, North-West Europe or South-Central Europe.
Let Z_i be an indicator of exposure for the ith patient, in our case Z_i is defined as:

$$Z_i = \begin{cases} 1 & \text{Region North-West Europe} \\ 0 & \text{Region South-Central Europe} \end{cases}$$

Let $\boldsymbol{X}_i$ be the row vector of observed covariates, the confounders, for the i-th patient. Then the propensity score

$$e(\boldsymbol{X}_i) = P(Z_i = 1|\boldsymbol{X}_i), \tag{3.14}$$

for the i-th patient is the probability of exposure given the observed covariates $\boldsymbol{X}_i$ [49].

Assuming a binary exposure variable, the propensity score is typically estimated for each individual using a logistic regression model

$$e(\boldsymbol{X}_i) = \frac{\exp(\boldsymbol{X}_i\beta)}{1 + \exp(\boldsymbol{X}_i\beta)}, \tag{3.15}$$

where β represents a vector of parameters to be estimated from the data [49].

A key theoretical property of the propensity score is described in Theorem 1.

Theorem 1 *Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is*

$$X \perp Z \mid e(X)$$

[38]

This property implies that, if we condition on the propensity score, the distribution of the observed covariates is the same in the treated and untreated groups. In other words, once we adjust for the propensity score, treatment assignment behaves as if it were randomised. So, this property allows for the creation of weighted samples that mimic randomised controlled trials. By creating these balanced samples, the confounding bias can be minimised, as the two groups become more comparable.

The variables $\boldsymbol{X}_i$ to include in the propensity score model are all baseline covariates that could confound the relationship between the exposure and the outcome, and covariates known to be associated only with the outcome [10]. When, according to the literature, one confounding variable is expected to be dependent on another confounding variable, interactions need to be included in the model.
It is important to note that the propensity score can only adjust for measured confounders [10].

IPTW
Using the propensity score defined in equation 3.14, the IPTW weights for the i-th individual are calculated as the inverse probability of being exposed :

$$W_i = \frac{Z_i}{e(X_i)} + \frac{(1 - Z_i)}{1 - e(X_i)} \tag{3.16}$$

After assigning the calculated IPTW weights and generating the pseudo-population, it is essential to assess whether balance has been achieved between the observed individual characteristics of the two groups. This is typically done by evaluating the standardised mean differences (SMDs) for all baseline covariates between the exposed and unexposed groups, both before and after weighting [10]. The SMD is defined as [4]:

$$SMD = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\left(S_T^2 + S_C^2\right)/2}}, \tag{3.17}$$

where $\bar{X}_T$ and $\bar{X}_C$ are the sample mean of the covariate in exposed and unexposed individuals, respectively, and S_T^2 and S_C^2 are the sample variance of the covariate in exposed and unexposed groups, respectively. The SMD formula for weighted observations is the same except that the weighted sample means and sample variances are used, which are given in:

$$\tilde{X}_T = \frac{\sum_{i\in\mathcal{T}} w_i X_i}{\sum_{i\in\mathcal{T}} w_i} \tag{3.18}$$

$$S_T^2 = \frac{1}{\sum_{i\in\mathcal{T}} w_i} \sum_{i\in\mathcal{T}} w_i \left(X_i - \widetilde{X}_T\right)^2 \tag{3.19}$$

Most researchers consider balance to be achieved when the absolute value of SMD is < 0.1. Guidelines indicate that 0.1 represents a reasonable cut-off point for acceptable standardised biases; larger standardised biases would indicate a too large difference between groups for reliable comparison [41].

Some individuals are likely to have a very high or low probability of being exposed. Taking the inverse of the propensity score can then lead to extreme weight values, which inflate the variance and confidence intervals of the effect estimate [10]. To deal with extreme weights, weight stabilisation or weight truncation can be used, which will be explained below.

When the original weights in equation 3.16 are used, the size of the pseudo-population is multiple times that of the original study population. To properly adjust for confounding, we aim to create a pseudo-population where the probability of receiving a certain exposure is the same for everyone, so this means that in the pseudo-population, assignment to an exposure does not depend on the confounders. To achieve this, the pseudo-population is created by applying weights that simulate a scenario where assignment to an exposure is random, with the same probability p for everyone, regardless of their confounders [19]. Thus, the IP weights are then $\frac{p}{e(X_i)}$ for the exposed and $\frac{1-p}{1-e(X_i)}$ for the unexposed. These weights are referred to as the stabilised weights. These weights also reduce the weights of exposed individuals with low propensity scores and unexposed individuals with high propensity scores [49]. By limiting the influence of extreme values, stabilised weights typically result in more precise estimates with lower variance [43]. The stabilised weights are calculated as in:

$$SW_i = \frac{pZ_i}{e(X_i)} + \frac{(1-p)(1-Z_i)}{1-e(X_i)}, \tag{3.20}$$

where p represents the probability of being assigned to an exposure without accounting for covariates. [49] With these stabilised weights, the size of the pseudo-population equals that of the original study population, illustrated in [49].

An additional strategy for reducing extreme weights is to set them to a less extreme value, called weight truncation or trimming. The weights are truncated by setting any values below the p-th percentile to the value of the p-th percentile, and any values above the $(100-p)$-th percentile to the value of the $(100-p)$-th percentile [11]. The 1st and 99th percentile can, for example, be used for this truncation [10].

3.3. Adjusted Kaplan-Meier estimator and weighted log-rank test

As mentioned before in Section 3.2, in a non-randomised clinical trial or observational study, the samples may be biased due to some confounding variables. Estimating a survival function using the Kaplan-Meier estimates may then be biased due to the unbalanced distribution of confounders. To overcome this problem, [48] has developed an adjusted Kaplan-Meier estimator using IPTW. Using this adjusted Kaplan-Meier estimator, one can provide comparable estimates when studying survival curves of two groups [48]. For comparing group differences of survival functions, a weighted log-rank test is also proposed by [48]. The formulas for the adjusted Kaplan-Meier estimator and weighted log-rank test will be given here. However, for details about the derivations and more information, we refer the reader to [48].

We use the formula for the standard Kaplan-Meier estimate as a starting point, given in equation 3.4. Using the IP weights W_i of equation 3.16, Z_i the indicator of the exposure, the weighted number of events at time u_i, d_j^w, and the weighted number of individuals still at risk at time u_i, r_i^w, are defined as:

$$d_j^w = \sum_{i:T_i=u_j} W_i \delta_i I(Z_i = 1) \tag{3.21}$$

$$r_j^w = \sum_{i:T_i \geq u_j} W_i I(Z_i = 1) \tag{3.22}$$

Using weighted formulas, the adjusted Kaplan-Meier estimate is defined as [48]:

$$\hat{S}^k(t) = \prod_{j:u_j \leq u} \left(1 - \frac{d_j^w}{r_j^w}\right), \tag{3.23}$$

The adjusted Kaplan-Meier estimate maximises a pseudo-likelihood function for survival data. Considering the case with only two groups, denoted by $Z_i = 0$ and $Z_i = 1$, the log-pseudo-likelihood of one group, where $Z_i = 1$, is defined as:

$$\sum_{i=1}^{n} \frac{Z_i}{p_i} \left[\delta_i \log\left(S_i\left(T_i - 0\right) - S_i\left(T_i\right)\right) + \left(1 - \delta_i\right) \log\left(S_i\left(T_i\right)\right)\right]$$

where $S_i(T_i - 0)$ includes but $S_i(T_i)$ excludes the probability of death exactly at T_i, and where $\frac{Z_i}{p_i}$ equals the weight W_i, applied only to the individuals in the group where $Z_i = 1$ [48]. For the derivation of this pseudo-likelihood function and the proof that the adjusted Kaplan-Meier estimate maximises this pseudo-likelihood function, we refer the reader to [48].

To provide the formula of the weighted log-rank test, the formula of the standard log-rank test statistic is used, given in equation 3.8. The weighted number of events and the weighted number of individuals at risk need to be combined into a pooled sample; therefore, the weights need to be adjusted in each group [48]. At time u_j, $j = 1, \ldots, m$, the weight is reassigned as $r_{j1} \times W_i / \sum_{i:Z_i=1} W_i$, for individual i in the group where $Z_i = 1$, and $r_{j0} \times W_i / \sum_{i:Z_i=0} W_i$ for individual i in the other group. In this way, the weights are proportional to the number of individuals at risk in each group [48]. Denote d_j^w and r_j^w as the weighted total number of individuals who had the event and were at risk, respectively. The weighted log-rank statistic under the null hypothesis is proposed as

$$X^2 = \frac{(G^w)^2}{Var(G^w)}, \tag{3.24}$$

where

$$G^w = \sum_{j=1}^{m} \left(d_{j1}^w - \frac{d_j^w r_{j1}^w}{r_j^w}\right) \tag{3.25}$$

and

$$\mathrm{Var}\left(G^w\right) = \sum_{j=1}^{m} \left\{ \frac{d_j\left(r_j - d_j\right)}{r_j\left(r_j - 1\right)} \sum_{i=1}^{r_j} \left[\left(\frac{r_{j0}^w}{r_j^w}\right)^2 W_i^2 Z_i + \left(\frac{r_{ji}^w}{r_j^w}\right)^2 W_i^2 \left(1 - r_i\right) \right] \right\} \tag{3.26}$$

3.4. Inverse Probability Weighted Cox Models

In the unweighted partial loglikelihood in the standard Cox model, given in equation 3.10, all patients contribute equally to estimating the regression coefficients. But this is not desirable when the data contain some confounding covariates. Section 3.2 describes how to account for those confounding covariates. For this situation, a weighted version of the Cox regression model is needed that includes patients of all subgroups but assigns them individual weights based on their subgroup affiliation [28]. The weights derived in the IPTW model have been used. Each weight reflects how much influence the corresponding individual has on the estimation of the regression coefficients [28].

An inverse probability weighted Cox model is fitted, similarly to the original Cox model explained in Section 3.1.4, by maximising a partial likelihood. But in the IP-weighted Cox model, this is the weighted partial log likelihood.
The probability that exactly the individual with index i had the event at time t from baseline is now approximately equal to [8]:

$$\left\{ \frac{e^{\mathbf{X}_i^T \hat{\boldsymbol{\beta}}}}{\sum_{j \in R(t)} \hat{w}_j e^{\mathbf{X}_j^T \hat{\boldsymbol{\beta}}}} \right\}^{\hat{w}_i}, i \in R(t),$$

The weighted partial log-likelihood for a Cox model is therefore defined as [7]:

$$\ell(\hat{\beta}) = \sum_{i=1}^{n} \hat{w}_i \delta_i \left(\mathbf{X}_i^T \hat{\boldsymbol{\beta}} - \log \left(\sum_{j \in R(T_i)} \hat{w}_j e^{\mathbf{X}_j^T \hat{\boldsymbol{\beta}}} \right) \right) \tag{3.27}$$

where

- $\exp(\beta)$ is the marginal hazard ratio associated with a one-unit increase in exposure $\boldsymbol{X}$, after adjusting for confounding and selection bias,
- The adjustment is achieved through the IP-weight $\hat{w}_i(t)$, which incorporates the effects of covariates used to control confounding,
- $\hat{\beta}$ is the pseudo-maximum likelihood estimator.

The vector $\boldsymbol{X}$ can now also, similar to the original Cox model, consist of other covariates that need adjustment. When those covariates were already added to the PS model, they were doubly adjusted for any residual confounding. The covariates for which the balance was not achieved after IPTW can then tried to be adjusted for any residual confounding again.

In the pseudo-population that is created during IPTW, some individuals contribute more than others to the estimation of the exposure effects, which could be interpreted as they appeared ”more than once”, although this isn't literally the case. Each individual is still a distinct observation in the dataset, so you don't lose statistical independence in the strict sense. However, because individuals are weighted differently, the resulting pseudo-population does not behave like a simple random sample anymore. As a result, standard error may be biased if standard (unweighted) methods are used. That is why it is important to use robust variance estimators when analysing IPTW-weighted data, as is used in the weighted Cox proportional hazards model.

4

Longitudinal trajectories of biochemical parameters using the latent class linear mixed model

This chapter addresses the first objective of this thesis: to determine and identify similarities of trajectories of relevant biochemical parameters, specifically sBA levels and platelet counts, in patients with PFIC2.

The dataset includes longitudinal laboratory measurements from PFIC patients, allowing us to explore how these biomarkers evolve over time. We hypothesise that the identification of patient subgroups by analysing longitudinal biochemical parameters may help better understand the disease and treat patients. In particular, distinguishing patients at a higher risk could support the development of targeted therapeutic interventions and enable earlier, more effective treatments.

To explore this, we focus on two parameters potentially indicative of disease severity: sBA levels and platelet counts. Their longitudinal trajectories are analysed using the latent class linear mixed model (LCLMM). This is a statistical method that identifies subpopulations (latent classes) based on similarities in individual progression patterns over time [17]. The LCLMM assumes a heterogeneous population composed of K latent classes, each defined by a distinct mean trajectory profile [34].

Before presenting the results in Section 4.3, Sections 4.1 and 4.2 explain the model. The first section will first focus on the (homogeneous) linear mixed model, and the second section focuses on the extension of this model, the heterogeneous linear mixed model, which is the LCLMM.

4.1. Homogeneous linear mixed model

The usual linear mixed model is a common approach for analysing longitudinal Gaussian outcomes over time and evaluating the impact of covariates [27].

Two key components of the linear mixed models are the fixed effects and the random effects:

- fixed effects describe how specific covariates influence the average longitudinal evolution
- random effects describe how specific regression coefficients deviate from the overall mean described by the fixed effects. The random effects also model the correlations in the repeated measurements.

In the linear mixed model, it is assumed that the vector of observations for the i-th individual, with n_i number of measurements, can be modelled as [45]

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, i = 1, \ldots, N \tag{4.1}$$

with $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ and $\mathbf{e}_i \sim \mathcal{N}(0, \sigma^2 I_{n_i})$ and where:

- $\mathbf{y}_i$ is an $n_i \times 1$ vector of longitudinal observations for individual i, where n_i is the number of measurements for individual i, at different time points.
- $\mathbf{X}_i$ is an $n_i \times p_1$ design matrix for the fixed effects
- $\boldsymbol{\beta}$ is a $p_1 \times 1$ vector of fixed effects
- $\mathbf{Z}_i$ is an $n_i \times q$ design matrix for the random effects
- $\mathbf{b}_i$ is an $q \times 1$ vector of random effects
- $\mathbf{D}$ is the variance-covariance matrix (symmetric positive definite) of the random effects
- $\mathbf{e}_i$ is an $n_i \times 1$ vector of random error terms with variance $\sigma^2 I_{n_i}$
- Furthermore, $\mathbf{b}_i$ and $\mathbf{e}_i$ are independent.

Each individual in the population has their own individual-specific mean response profile over time, represented by the sum $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$. The term $\mathbf{X}_i\boldsymbol{\beta}$ corresponds to the fixed effects, which describe the population-average trajectory, while the term $\mathbf{Z}_i\mathbf{b}_i$ captures individual-specific deviations from this average trajectory through the random effects.

4.2. Heterogeneous linear mixed model - LCLMM

The homogeneous linear mixed model assumes that the population of N individuals is described at the population level by a single average trajectory, given by $\mathbf{X}_i\boldsymbol{\beta}$ [34]. This expression corresponds to the unconditional expectation of the response vector $\mathbf{y}_i$. The heterogeneous linear mixed model extends this homogeneous model, in that each individual's longitudinal data is modelled as a mixture of K linear mixed models, where each one of the K linear mixed models corresponds to one latent class. This extended linear mixed model, called the heterogeneous linear mixed model or LCLMM, now allows one to identify distinct trajectories of the marker and group individuals based on these different trajectories [33].

Each individual belongs to one latent class. For the i-th individual, let c_{ik} be the indicator variable that denotes whether individual i belongs to class k, for $i = 1, \ldots, N$ and $k = 1, \ldots, K$, i.e.

$$c_{ik} = \begin{cases} 1 & \text{if individual } i \text{ is a member of class } k, \\ 0 & \text{if individual } i \text{ is not a member of class } k. \end{cases}$$

The probabilities π_{ik} of latent class membership explained according to covariates $\boldsymbol{\nu}_i$ are given by the multinomial logistic regression model [2] [17]:

$$\pi_{ik} = \mathbb{P}\left(c_{ik} = 1 \mid \boldsymbol{\nu}_i\right) = \frac{\exp\left(\xi_{0k} + \boldsymbol{\nu}_i'\xi_{1k}\right)}{\sum_{j=1}^{K} \exp\left(\xi_{0j} + \boldsymbol{\nu}_i'\xi_{1j}\right)}, \tag{4.2}$$

for $k = 1, \ldots, K$ and $i = 1, \ldots, N$ and where

- $\boldsymbol{\nu}_i$ are variables related to the membership of the class for the individual i
- ξ_{0k} is the intercept for class k and ξ_{1k} is the vector of class membership parameters for class k. For identifiability, the constraints $\xi_{0K} = 0$ and $\xi_{1K} = 0$ are included.

The K mean profiles are modelled over time and covariates using latent class-specific mixed models. Within each class, the profiles are modelled by a standard linear mixed model [34].

Furthermore, given that individual i is in class k, the general formulation of the LCLMM for individual i is defined as:

$$\mathbf{y}_i = \mathbf{X}_{2i}\boldsymbol{\beta} + \mathbf{X}_{3i}\boldsymbol{\gamma}_k + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{4.3}$$

where:

- $\mathbf{X}_{2i}$ is an $n_i \times p_1$ design matrix for the fixed effects (common over all classes)
- $\boldsymbol{\beta}$ is an $p_1 \times 1$ vector of fixed effects (common over all classes)
- $\mathbf{X}_{3i}$ is an $n_i \times p_2$ design matrix for the class-specific fixed effects
- $\boldsymbol{\gamma}_k$ is an $p_2 \times 1$ vector of class-specific fixed effects for class k

- $\mathbf{Z}_i$ is an $n_i \times q$ design matrix for the random effects
- $\mathbf{b}_i$ is an $q \times 1$ vector of random effects. These distributions are now class-specific, that is, for individual i in class k: $b_{ik} \sim \mathcal{N}(0, \mathbf{D}_k)$ where $\mathbf{D}_k = \omega_k^2 D$ with $\mathbf{D}$ an unspecified variance-covariance matrix and ω_k a proportional coefficient for class k which adjusts the class-specific intensity of individual variability. For identifiablity, $\omega_K = 1$ [33][34].

The design matrix for the fixed effects from equation 4.1 is split in the design matrices $\mathbf{X}_{2i}$ and $\mathbf{X}_{3i}$, but also in an extra matrix $\mathbf{X}_{1i}$ that contains covariates for the class-membership part [33]. So these are the covariates predicting latent class membership via multinomial logistic regression. The equation for the probabilities π_{ik} in equation 4.2 can then be changed to

$$\pi_{ik} = P\left(c_{ik} = 1 \mid \mathbf{X}_{1i}\right) = \frac{e^{\xi_{0k} + \mathbf{X}_{1i}^\top \xi_{1k}}}{\sum_{j=1}^{K} e^{\xi_{0j} + \mathbf{X}_{1i}^\top \xi_{1j}}} \tag{4.4}$$

It is important to note that the parameters in $\boldsymbol{\beta}$ apply to all individuals, as they are linked to the values of the corresponding column in the design matrix $\mathbf{X}_{2i}$. In contrast, the class-specific parameters γ_k differ across latent classes and capture the unique characteristics of each class [16].

Given that individual i is in class k, $c_{ik} = 1$ and the parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}_k, \mathbf{D}, \sigma^2$ the distribution of the observations is then given by

$$\mathbf{y}_i \mid (c_{ik} = 1, \boldsymbol{\beta}, \boldsymbol{\gamma}_k, \mathbf{D}, \sigma^2) \sim \mathcal{N}\left(\mathbb{E}[\mathbf{y}_i \mid c_{ik} = 1, \boldsymbol{\beta}, \boldsymbol{\gamma}_k, \mathbf{D}, \sigma^2], \mathsf{Var}[\mathbf{y}_i \mid c_{ik} = 1, \boldsymbol{\beta}, \boldsymbol{\gamma}_k, \mathbf{D}, \sigma^2]\right) \tag{4.5}$$

with

$$\begin{aligned}
\mathbb{E}[\mathbf{y}_i \mid c_{ik} = 1, \boldsymbol{\beta}, \boldsymbol{\gamma}_k, \mathbf{D}, \sigma^2] &= \mathbf{X}_{2i}^\top \boldsymbol{\beta} + \mathbf{X}_{3i}^\top \boldsymbol{\gamma}_k \\
\mathsf{Var}[\mathbf{y}_i \mid c_{ik} = 1, \boldsymbol{\beta}, \boldsymbol{\gamma}_k, \mathbf{D}, \sigma^2] &= \mathbf{Z}_i^\top \mathbf{D}_k \mathbf{Z}_i + \sigma^2 I_{n_i},
\end{aligned}$$

which leads to the marginal distribution of y_i being a finite mixture of K distributions:

$$\mathbf{y}_i \mid \pi_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{D}, \sigma^2 \sim \sum_{k=1}^{K} \pi_{ik} \mathcal{N}\left(\mathbf{X}_{2i}^\top \boldsymbol{\beta} + \mathbf{X}_{3i}^\top \boldsymbol{\gamma}_k, \mathbf{Z}_i^\top \mathbf{D}_k \mathbf{Z}_i + \sigma^2 I_{n_i}\right) \tag{4.6}$$

[17]

4.2.1. Estimation

The LCLMM is estimated using maximum likelihood estimation. Let $\boldsymbol{\theta}_K$ denote the complete parameter vector, with estimation conducted for a fixed number K of latent classes. The value of K is specified by the statistician a priori, but to choose the optimal number of classes, some information criteria have been used. More details about the selection will be explained at the end of this section.
The individual contribution to the likelihood of an LCLMM is:

$$\begin{aligned}
L_i(\boldsymbol{\theta}_K \mid \mathbf{y}_i) &= \sum_{k=1}^{K} \pi_{ik} \phi_{ik}(\mathbf{y}_i \mid \boldsymbol{\theta}_K) \\
&= \sum_{k=1}^{K} \pi_{ik} \mathcal{N}\left(\mathbf{X}_{2i}^\top \boldsymbol{\beta} + \mathbf{X}_{3i}^\top \boldsymbol{\gamma}_k, \mathbf{Z}_i^\top \mathbf{D}_k \mathbf{Z}_i + \sigma^2 I_{n_i}\right)
\end{aligned} \tag{4.7}$$

where π_{ik} is given in equation 4.4 and ϕ_{ik} is the density function of a multivariate normal distribution. The log-likelihood of the model,

$$l_i(\boldsymbol{\theta}_K) = \sum_{i=1}^{N} \log(L_i(\boldsymbol{\theta}_K \mid \mathbf{y}_i)),$$

can be maximised using different algorithms, like the EM algorithm. But in the R package used in this research, the $lcmm$ package, the extended Marquardt algorithm is used [34]. When the number of latent classes K is fixed, the algorithm simultaneously estimates two types of parameters:

- The trajectory parameters that describe how each class evolves over time (e.g., intercepts, slopes, and other effects within each class)
- The probability that an individual belongs to a particular latent class.

To choose the optimal number of classes, one has to estimate models with different fixed numbers of latent classes and select the best model according to some criterion [33]. The log-likelihood, the Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the posterior proportion of each class have been used for the estimation process [34].

In model selection, a higher log-likelihood indicates a better fit to the data. The AIC, where lower values indicate a better fit, is computed as $AIC = -2L + 2P$, where P is the number of parameters. The BIC, the lower the better, is computed as $BIC = -2L + P\log(N)$, where N is the number of individuals.

After model estimation, individuals are assigned to the class for which they have the highest posterior probability. For an individual i in latent class k the posterior probability of membership is computed using the Bayes theorem and results as follows [34] [33]:

$$\hat{\pi}_{ik}^{(\mathbf{y})} = P(c_{ik} = 1 \mid \mathbf{X}_{1i}, \mathbf{X}_{2i}, \mathbf{y}_i, \hat{\boldsymbol{\theta}}_K) = \frac{\pi_{ik}\phi_{ik}(\mathbf{y}_i \mid c_{ik} = 1, \hat{\boldsymbol{\theta}}_K)}{\sum_{j=1}^{K} \pi_{ij}\phi_{ij}(\mathbf{y}_i \mid c_{ij} = 1, \hat{\boldsymbol{\theta}}_K)}, \tag{4.8}$$

where $\phi_{ik}(Y_i \mid c_{ik} = 1, \hat{\theta}_K)$ is the density function of a multivariate normal distribution derived in equations 4.5 and 4.6.
This probability tells us how likely it is that an individual i belongs to a certain class given their observed data.

4.3. Results identified trajectories of sBA levels and platelet counts

As explained at the beginning of this chapter, this part of the thesis aims to determine and identify similarities of trajectories of the biochemical parameters, sBA levels and platelet counts, in patients with PFIC. Actually, the focus on this part of the study is on patients with PFIC2, since this is the largest group in the dataset. This section gives the results of the LCLMM for identifying the trajectories.

The dataset used in this thesis, introduced in Chapter 2, consists, among others, of longitudinal laboratory measurements of the biochemical parameters sBA levels and platelet counts. The trajectories of these markers are modelled separately using the LCLMM model according to the age covariate.

To perform the analysis, a selection of patients and visits was made, the flowchart in Figure 4.1 shows the exclusion criteria. This gives a selection of 167 patients and 946 measurements, with a median of 4 (IQR: 2-8) measurements per patient.

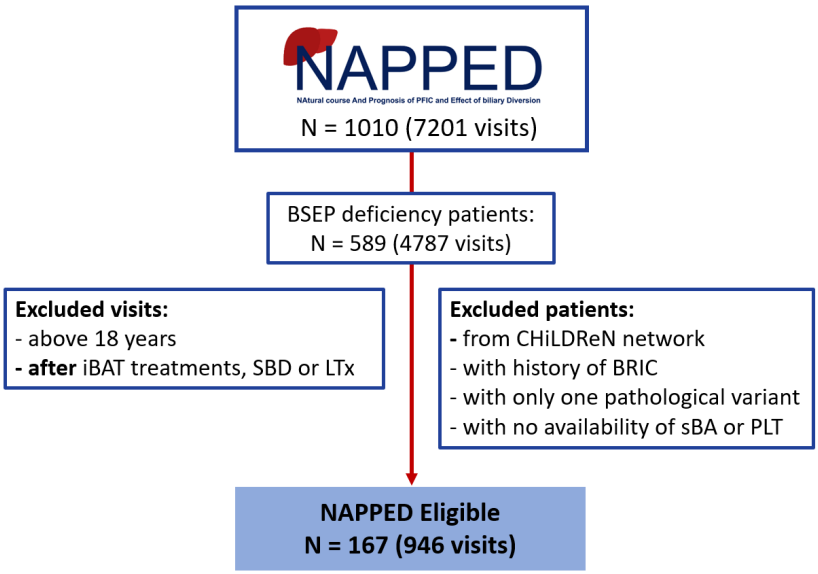


Figure 4.1: Flowchart of the exclusion criteria for the selection of patients for the LCLMM analysis

In the latent class mixed model, we have added time dependency to the model using nonlinear terms of age as fixed and random effects. Spline functions are used to include this nonlinearity. If we were to include the age effect as a main effect, it would give the assumption that the effect of age is linear. To relax this assumption, nonlinear terms have to be included. Using splines allows the age effect to be modelled more flexibly [37].

Splines are piecewise polynomials. This means that they are polynomials within intervals of X that are connected across different intervals of X [18]. The limits of these intervals are defined by the knots of the spline, with two boundary knots and a number of internal knots [37]. A cubic spline is a function defined by cubic polynomials that are spliced together at knot locations [13]. They can be made to be smooth at the knots by forcing the first and second derivatives of the function to agree at the knots [18]. Natural cubic splines have the additional constraints of having a second derivative of zero at the boundaries [13].

The locations of the knots have to be specified in advance. Placing knots at fixed quantiles of the marginal (empirical) distribution of the predictor is a good approach in most datasets, according to literature [18]. This thesis consists of small data samples, and therefore, according to [18], the natural cubic spline function with three knots (one interior, and two boundary knots) has been used in this research, and the recommended equally spaced quantiles for these knots are $0.10, 0.5, 0.90$.

The spline functions can be expressed as a linear combination of basis functions, $N_l(t)$, and weights β_l: $f(t) = \beta_0 + \sum_{l=1}^{m} \beta_l N_l(t)$ [13], with m the number of knots. The basis functions $N_l(t)$ can be numerically

determined given the boundaries, interior knot locations, and the continuity and derivative constraints. This is done using the $ns()$ function in R.

The following latent class mixed model is considered, where k denotes the class, i denotes the individual, and j denotes the repeated measurement. The basis spline function for age is denoted as $\sum_{l=1}^{m} N_k(age)$. Equations 4.9 and 4.10 denote the LCLMMs for platelet counts (PLT) and sBA levels, respectively. Given that individual i is in class k the model is defined as:

$$PLT_{ij} = \beta_0 + \sum_{l=1}^{3} \beta_l N_l(age_{ij}) + b_{i0} + \sum_{l=1}^{3} b_{il} N_l(age_{ij}) + \varepsilon_{ij}, \tag{4.9}$$

$$(\log sBA)_{ij} = \beta_0 + \sum_{l=1}^{3} \beta_l N_l(age_{ij}) + b_{i0} + \sum_{l=1}^{3} b_{il} N_l(age_{ij}) + \varepsilon_{ij}, \tag{4.10}$$

where $b_i \sim \mathcal{N}(0, D)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, using the same terminology as in the general LCLMM given in equation 4.3, also for the fixed and random effects β and b_i. The term $\sum_{l=1}^{3} N_l(age_{ij})$ denotes the basis for a natural cubic spline with 3 knots. In both the LCLMMs for PLT and sBA, the fixed and random effects are modelled as 'non-linear age effects using splines'. This means that age is deemed as an overall fixed effect that models the 'effect of shared age' between all individuals. The mixture of the spline function of age means that the spline function of age is considered for the class-specific fixed effects. Here, age is used to define separate average trajectories for each latent class. It's fixed within each class but differs across classes.
The random part is $b_{i0} + \sum_{l=1}^{3} b_{il} N_l(age_{ij})$. This random effects part models individual deviations from their latent class' trajectory, which means that each person can follow their age curve. To improve the normality of the residuals, we have log-transformed the sBA values before adding them to the latent class mixed model.

As the number of classes is unknown, models with $K = 1, 2, 3$ classes are evaluated. The $hlme$ function of the R package $lcmm$ is used to model the latent class mixed models [34]. For the estimation of the model with more than one class ($K > 1$), the initial values must be given since it is an iterative estimation. The initial values are crucial for the convergence of the algorithm to the true maximum [34]. We have used the gridsearch function, which is used to run an automatic grid search. The procedure involves running the estimation function $hlme$ for a maximum of m iterations from B randomly chosen vectors of initial values. The initial values from the maximum likelihood estimates of the 1-class model are used for this. The set of parameters that provides the best log-likelihood after m iterations is then used as the initial value for the final parameter estimation [34].

The class-specific predictions can be computed after the latent class mixed model is fitted, using the parameter estimates from the model. The predicted mean trajectory of the longitudinal outcome variables, the biochemical parameters PLT and sBA in our case, according to a hypothetical profile of covariates, age in our case, can be computed and presented using the functions $'predictY'$ and the plot function applied on $'predictY'$ objects of the $lclmm$ package in R [34]. The uncertainty around the predicted trajectories is assessed by this R function by approximating the posterior prediction distribution using a Monte Carlo method [34]. The 2.5%, 50% and 97.5% percentiles provide the mean prediction and its 95% simulated prediction bands.

The identified classes for the sBA and PLT latent class mixed models across each value of K are shown in Figures 4.2 and 4.3. The percentages of individuals assigned to each class per model are given in Table 4.1. To determine the optimal number of classes, the information criteria, AIC and BIC, were calculated for each model, as shown in Table 4.1. Detailed information about these information criteria is given in the previous section. In theory, the model with the lowest value of the information criterion used should be selected. However, in practice, additional factors such as interpretability and class size must also be considered. Therefore, we selected the model in which all latent classes comprised more than 10% of the study population. This approach avoids selecting models with extremely small classes — for instance, one model included a class representing only 2% of participants. Such small classes typically add little value to the model, as they may reflect random variation or outliers rather than meaningful subgroups, and are often difficult to interpret.

The results of the LCLMM for sBA indicate that when 3 classes are estimated, one class only consists of 2.40% of the study population, which means that this model is not considered. While the log-likelihood, AIC, and BIC values improve from 1 to 3 classes, the 2-class model offers the best balance between model fit and class stability. Therefore, the 2-class solution is considered the most appropriate for the sBA model.

The results of the LCLMM for PLT indicate that the AIC remains relatively stable across models with different numbers of classes. The log-likelihood improves from the 1-class to the 3-class model, suggesting a better fit with more classes. However, the BIC increases from 1 to 3 classes. These mixed results make it unclear which model provides the best overall fit. While the 1-class model may oversimplify the data and fail to capture heterogeneity between individuals, the 3-class model does not offer a substantial improvement over the 2-class model. Therefore, the 2-class model is preferred for the PLT data.

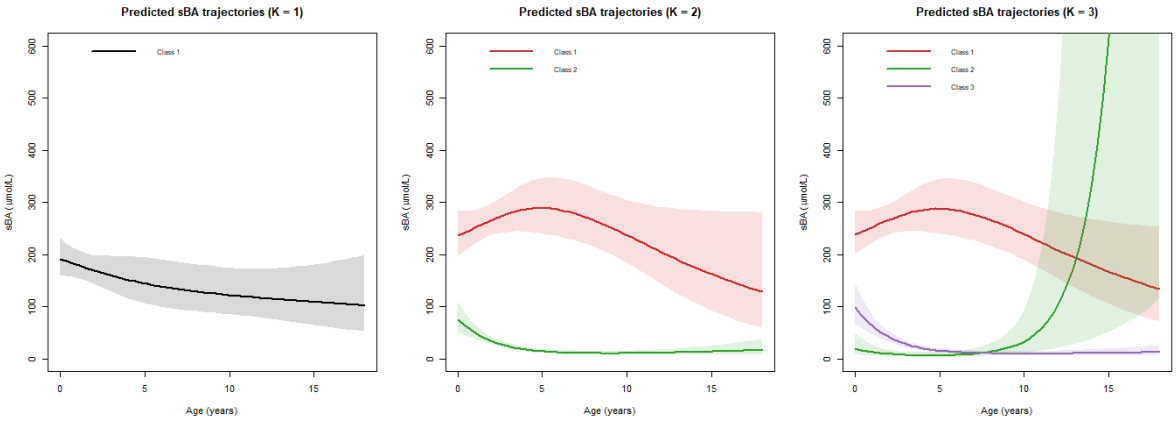


Figure 4.2: Predicted mean sBA trajectories of the LCLMM with $K = 1$, $K = 2$, and $K = 3$ classes, respectively.

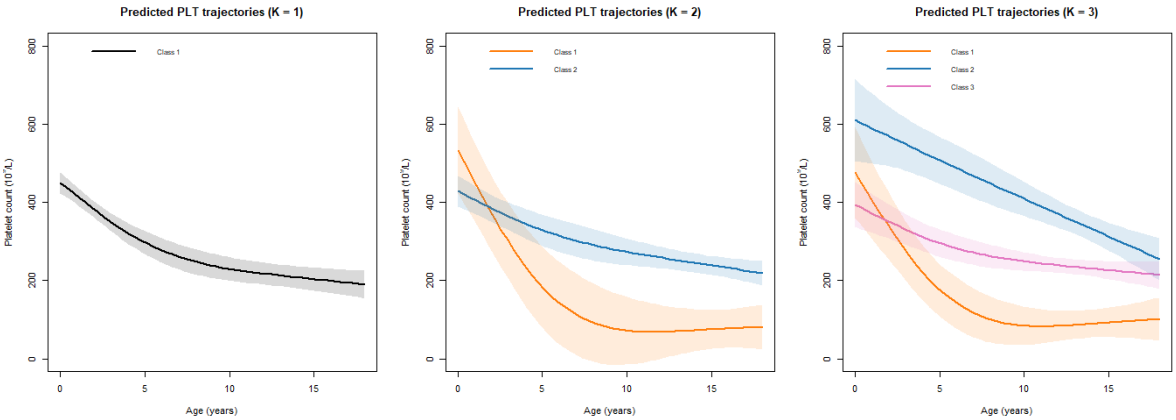


Figure 4.3: Predicted mean PLT trajectories of the LCLMM with $K = 1$, $K = 2$, and $K = 3$ classes, respectively.

Table 4.1: Information criteria of the results of the latent class mixed models for sBA and PLT

	K	loglik	AIC	BIC	%Class1	%Class2	%Class3
sBA model	1	-956.16	1932.31	1963.49	100.0		
sBA model	2	-921.39	1870.78	1914.43	84.43	15.57	
sBA model	3	-914.35	1864.71	1920.83	83.8	2.40	13.77
PLT model	1	-5299.76	10619.52	10650.70	100.0		
PLT model	2	-5295.94	10619.89	10663.54	13.17	86.83	
PLT model	3	-5291.69	10619.37	10675.50	16.77	14.97	68.26

Therefore, we can identify two distinct trajectory classes for the sBA and also two distinct trajectory classes for the PLT model. Figure 4.4 shows the observed sBA and PLT measurements before and after the patients have been identified in classes by the latent class mixed model. The dots represent the observed measurements. Lines connect observations from the same individuals. The predicted mean trajectories of both sBA and PLT are added to these observed measurements plots and shown in Figure 4.5. The bold solid lines indicate the predicted mean trajectories of the longitudinal sBA and PLT parameters, and the shaded areas indicate their corresponding 95% prediction bands.

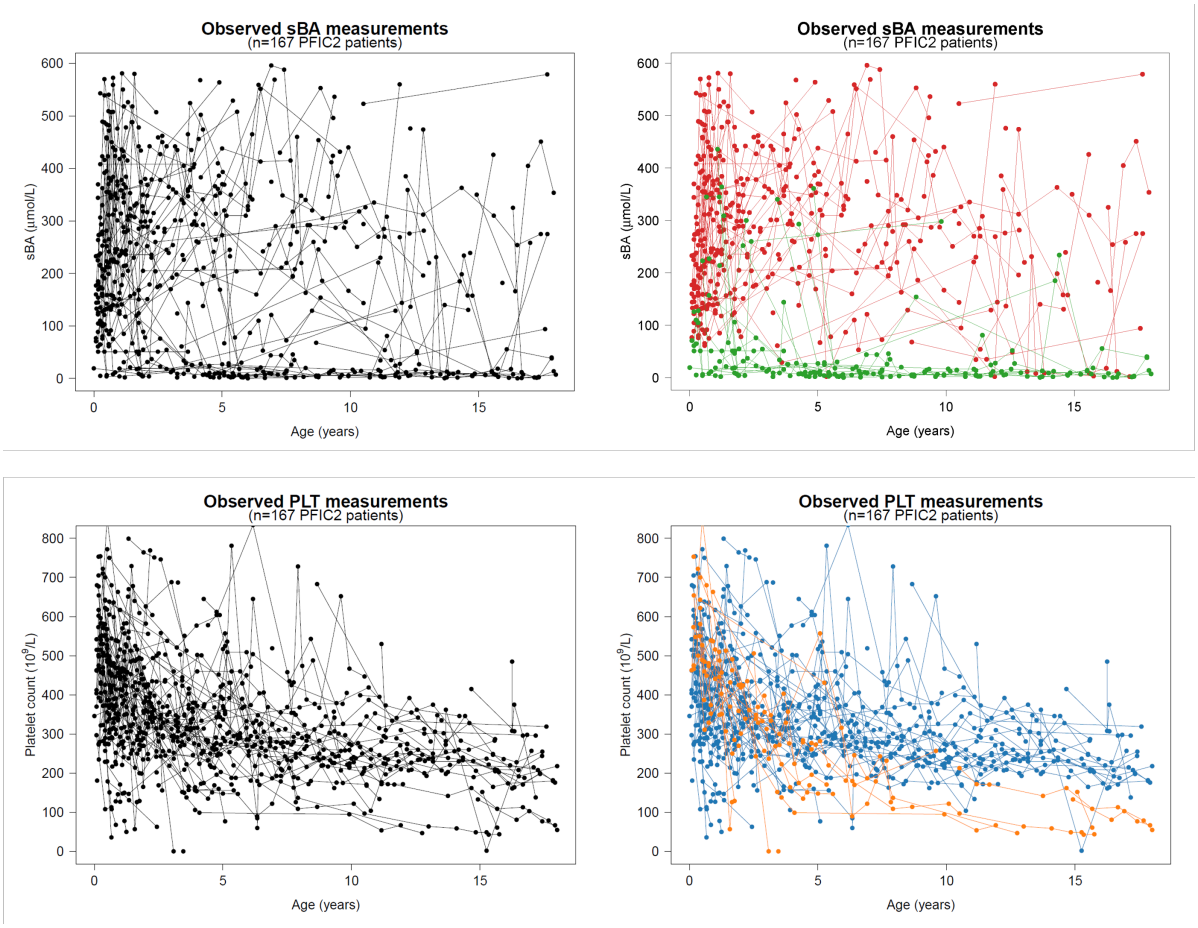


Figure 4.4: The two left plots show the observed sBA and PLT measurements of the selected PFIC2 patients. The plots on the right indicate the different classes that had been identified by the latent class mixed models.

Characteristics of the sBA and PLT trajectories stratified by class are given in Table 4.3. We will focus on some interesting observations from these results:

The results show that the Classes sBA1 and PLT1 contain more patients compared to the Classes sBA2 and PLT2, respectively.
The follow-up duration characteristic in Table 4.3 shows that Class sBA2 has been followed over a longer time compared to the patients in Class sBA1, this is probably because the patients in Class sBA1 experienced more first events.
Another interesting observation is that almost half of the patients in Class PLT2 experienced a liver transplantation at the end of their follow-up.

Analysing the results from a clinical perspective gives the speculation that the patients in the Class PLT2 have a more progressive course in platelet counts, which could perhaps identify patients with a more rapid development of overactive spleen (hypersplenism).
Another interesting result from a clinical perspective is Class sBA2, which is a rather unexpected observation for patients with PFIC2, since the values of the sBA levels are relatively low from a very low age.

Therefore, a detailed analysis is performed on two patients in this class to be able to provide a clearer explanation to clinicians that these patients do exist with PFIC2. Figure 4.6 gives the observed sBA measurements of two patients of Class sBA2. Both patients are BSEP2 patients, which is a specific form of the PFIC2 disease. From the observed measurement of patient 1, this patient seems to have had no high sBA levels throughout childhood, which could indicate that the patient did not suffer from severe cholestasis. Patient 2 had some high levels of sBA, with low levels in between. This seems to qualify as an episodic form of PFIC, but this is not indicated as such in the dataset. Thus, the patterns indicated by the latent class mixed model allow for better identification of subgroups, including even quality control, as well as better identification of patients with an episodic phenotype.

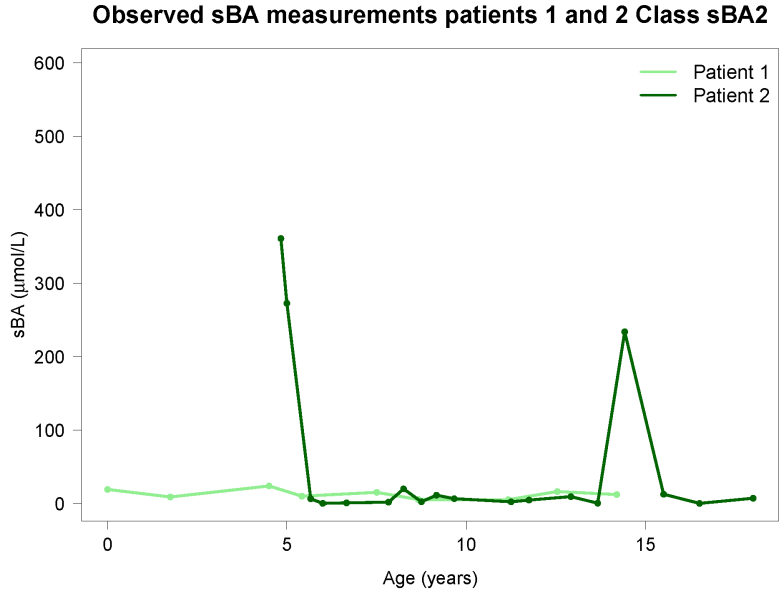


Figure 4.6: The observed measurements of 2 patients of the Class sBA2.

Table 4.2 shows the cross-table of the distribution of individuals across the classes. Only one patient falls into both Class sBA2 and PLT2, which highlights a clear distinction between these two classifications.

Table 4.2: Cross-table distributions classes

	Class PLT1	**Class PLT2**
Class sBA1	120	21
Class sBA2	25	1

The results show that there are distinct longitudinal sBA and PLT patterns identified in patients with PFIC2. These patterns reveal substantial heterogeneity in the course of laboratory parameters over time. This offers potential for trajectory-specific management strategies that could improve patient care and outcomes for patients with PFIC2.

For the interested reader, the code to reproduce these LCLMM results is available in the Github repository [1].

[1] https://github.com/paulinexhuisman/Code-Master-Thesis-Pauline-Analysis-of-PFIC-data

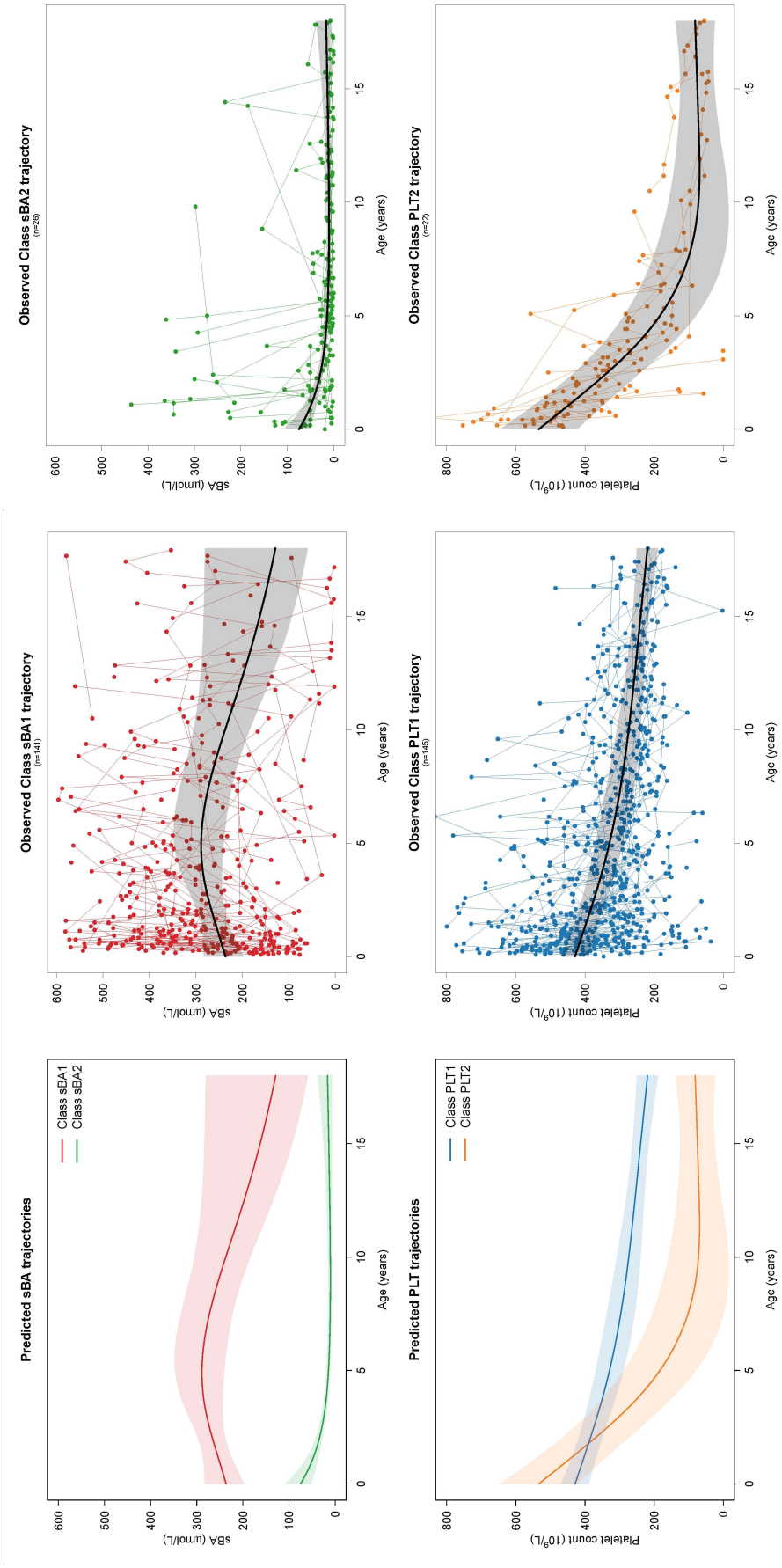


Figure 4.5: The final results of the LCLMMs for sBA and PLT. The observed measurements are shown, and the predicted mean trajectories with the 95% prediction bands.

Table 4.3: Characteristics of the sBA and PLT trajectories. Data are presented as *n*(%) or median (Q1-Q3).

Characteristics		**Class sBA1**	**Class sBA2**	**Class PLT1**	**Class PLT2**
Number of patients		141 (84.4)	26 (15.6)	145 (86.8)	22 (13.2)
Number of patients with first event	SBD	40 (28.4)	2 (7.69)	40 (27.6)	2 (9.09)
	LTx	44 (31.2)	3 (11.5)	37 (25.5)	10 (45.5)
	Death	3 (2.13)	0 (0)	1 (6.90)	2 (9.09)
Sex	Male	61 (43.3)	16 (61.5)	71 (49.0)	6 (27.3)
	Female	78 (55.3)	10 (38.5)	72 (50.0)	16 (72.7)
	Unknown	2 (1.42)	0 (0)	2 (1.38)	0 (0)
Follow-up duration, in years		1.25 (0.17-3.71)	6.29 (3.27-13.02)	1.34 (0.17-5.21)	2.67 (1.59-6.83)
BSEP category	BSEP1	42 (29.8)	8 (30.8)	44 (30.3)	6 (27.3)
	BSEP2	62 (44.0)	16 (61.5)	71 (49.0)	7 (31.8)
	BSEP3	37 (26.2)	2 (7.69)	30 (20.7)	9 (40.9)
Number of total measurements		681 (77.9)	193 (22.1)	723 (82.7)	151 (17.3)
Number of measurements per patient		2.00 (1.00-4.00)	4.50 (3.00-9.75)	3.00 (2.00-7.00)	6.00 (4.00-8.75)
Age across total measurements, in years		2.33 (0.75-7.06)	5.82 (2.40-11.17)	4.17 (1.25-8.92)	3.16 (1.25-6.67)

5

Comparison of survival time distributions using IPTW

This chapter presents the results of the comparison of survival time distributions between the two regional groups, North-West Europe and South-Central Europe, using IPTW. In Chapter 3, the methodology for performing this analysis is explained. To be able to perform a valid comparison, we need to have a balance of the characteristics, using IPTW. Using this technique helps by creating a hypothetical randomised trial which is perfectly balanced between the two groups. This hypothetical randomised trial is called the target trial. The first section of this chapter explains more about this target trial and how the target trial is emulated, which leads to the final dataset used in the analysis. Sections 5.2 and 5.3 give the mathematical notation and characteristics of the selected dataset. Following, the process and results of the IPTW method are given in Section 5.4. After which the results of the weighted Kaplan-Meier and weighted Cox regression model are presented, to give the final result of the second objective of this thesis, in Section 5.5. Finally, we tested the hypothesis of no difference between regional groups using another (type of) test, the permutation test.

For the interested reader, the code to reproduce the results explained in this chapter of the comparison of survival time distributions using IPTW is available in the Github repository [1].

[1] https://github.com/paulinexhuisman/Code-Master-Thesis-Pauline-Analysis-of-PFIC-data

5.1. Target Trial

The data used in this thesis are the data collected from the NAPPED database. As mentioned before, the second objective of this thesis is to perform a comparison of the survival until the first event of liver transplantation, SBD or death of the PFIC1 & PFIC2 patients in the two divided regions in Europe.
The countries in Europe have been divided into the following two regions, also visible in Figure 5.1, where the red countries indicate North-West Europe and the blue ones South-Central Europe:

North-West Europe	South-Central Europe
Denmark	Albania
Estonia	Cyprus
Finland	Greece
Iceland	Italy
Norway	Portugal
Sweden	Spain
Belgium	Austria
France	Bulgaria
Ireland	Croatia
Luxemburg	Czech Republic
Netherlands	Germany
United Kingdom	Hungary
	Latvia
	Poland
	Romania
	Russia
	Serbia
	Slovakia
	Slovenia
	Switzerland
	Ukraine

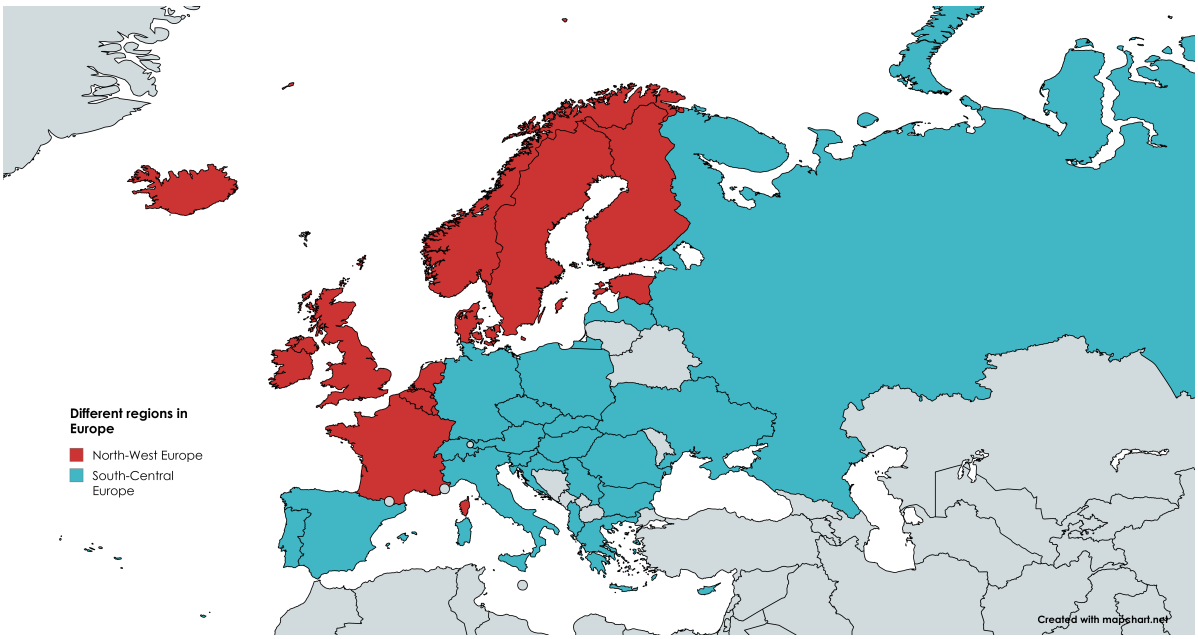


Figure 5.1: Division of Europe into the 2 different regions

The comparison is made between Western and Northern Europe versus Central and Southern Europe.

These regions are grouped to ensure approximately balanced group sizes of the patients in these regions, which is needed to perform a valid comparison and supports stable estimation in the IPTW procedure.

To perform a valid comparison between these two created regional groups, we first want to create a balance of the characteristics of the groups. We would like to have a hypothetical randomised trial which is perfectly balanced between the two groups, which is called the target trial. By emulating the target trial, we aim to approximate the effect estimates that such a trial would have produced. There are a few key components of the target trial protocol that are important for our analysis: eligibility criteria and the follow-up period (the choice of the time zero of follow-up) [20]. Section 5.1.1 gives a selection of the data using eligibility criteria, and Section 5.1.2 explains how to determine the follow-up period.

5.1.1. Data selection

The observational analysis should apply the same eligibility criteria used in the target trial. These are criteria which only include patients who are critically ill, and include the patients that we want to focus the research on:

1. The patient's age must be between 1 and 18 years.
2. The patient must be diagnosed after the year 1990.
3. The sBA level concentration must be $\geq 3 \times ULN_{sBA}$.
4. The alanine transaminase (ALT) concentration must be $\leq 15 \times ULN_{ALT}$.
5. The patient must have clinical genetic confirmation of PFIC1 or PFIC2.
6. The patient did not undergo a liver transplant or SBD on or before the index time date.

By applying these criteria to the data, we had to make some assumptions. In the original data, there were a lot of missing values in the variables ULN_{sBA} and ULN_{ALT}. So we have chosen that for missing values of all ULN values, the most frequent value of the specific ULN within the same hospital is used. When no ULN value for a hospital is given, the most frequent value of the ULN within the country is used for the remaining missing $ULNs$. Figure 5.2 shows the flow chart for the selection of the database.

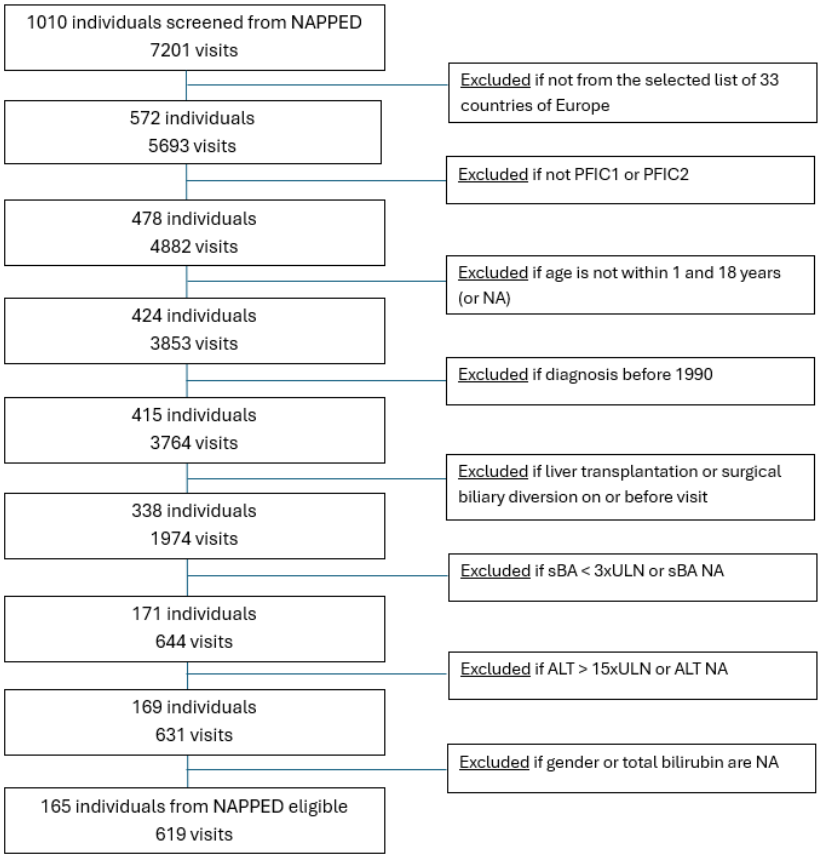


Figure 5.2: Selection of the NAPPED database using the selection criteria

All visits of patients that meet these criteria are called eligible.

5.1.2. Index time

A crucial component of the emulation of the target trial is the determination of the start of follow-up in the observational data, which is called the index time, time zero or baseline. The eligibility criteria must be met at that specific time [20]. From the date of this index time, the follow-up starts. The follow-up times are calculated per individual as the time from the selected index time to the first clinical event that occurs, or until the last day of follow-up. The different clinical events we take as endpoints are death, liver transplantation, and SBD.

The eligibility criteria can be met at many different times for the same individual, this is presented schematically in Figure 5.3. The blue bar represents one individual from the EU North-West region, and the green bar represents one individual from the EU South-Central region. The visits of an individual are represented by a diamond. An eligible visit (yellow) means that the individual has fulfilled all the inclusion and exclusion criteria at that time.
When an individual has multiple eligible visits, we have to decide on how to choose the time zero. The goal is that the two data cohorts are as similar as possible. We will try different index times to see if there is any difference in the results in a sensitivity analysis. The options we will investigate are selecting a random eligible visit per individual, selecting the first eligible time per individual, or selecting the last eligible time per individual. In Figure 5.3, selecting the first index time and random index time of all eligible visits is schematically shown.

We will perform a sensitivity analysis for these different index times to determine the degree to which this selection procedure affected study outcomes. The results of the sensitivity analysis are given in Section 5.5.1.
We take the first eligible visit as the selected index time as the standard index time in all the plots

given throughout this thesis, except as indicated, because using this index time the events occur with maximum possible follow-up.

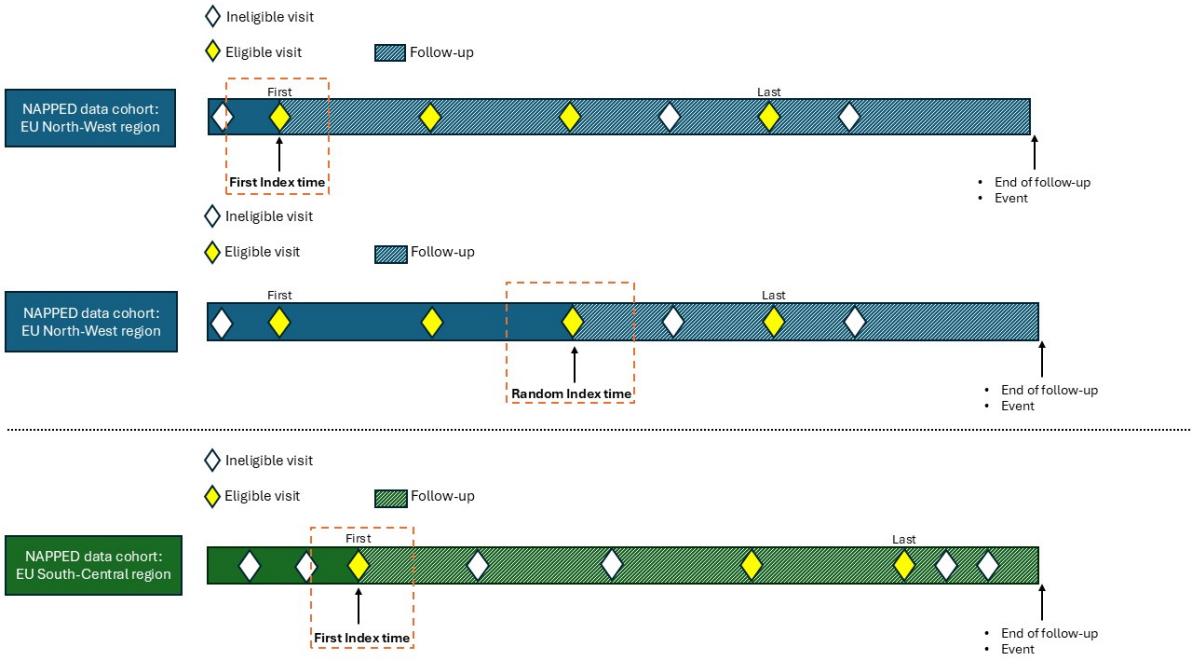


Figure 5.3: Schematic visualisation of the selection of the index time. The blue bar represents one individual from the EU North-West region with their ineligible and eligible visits, indicated with diamonds. And the green bar represents one individual from the South-Central region. This visualisation shows the difference in selecting the first eligible visit as the index time or a random eligible visit.

5.2. Mathematical notation for the selected dataset

The dataset that is created after applying the eligibility criteria and selecting the index times (baseline) no longer has the longitudinal format. The dataset now consists only of the baseline variables of each eligible patient. This dataset consists of the covariates which are chosen for a clinical reason, because these are the covariates for which we want to assess the balance. This newly created data is used in the following part of this research and is described mathematically as follows.

Denote the selected dataset by S as $S = \{(\mathbf{B}_i, \mathbf{C}_i, t_i, e_i)\}_{i=1}^{n}$, where the total number of patients is $n = 165$. Each i-th patient has:

- $\mathbf{B}_i \in \mathbb{R}^{n_{num}}$ represents the $n_{num} = 4$ numerical covariate features at baseline:

$$\mathbf{B}_i = \begin{pmatrix} \text{Age at baseline} \\ \text{sBA at baseline} \\ \text{ALT at baseline} \\ \text{Total Bilirubin at baseline} \end{pmatrix}$$

- $\mathbf{C}_i \in \mathbb{Z}^{n_{cat}}$ represent the $n_{cat} = 2$ categorical covariate features:

$$\mathbf{C}_i = \begin{pmatrix} \text{Sex} \\ \text{Genetic severity based PFIC type} \end{pmatrix}$$

- $t_i \in \mathbb{R}$ represents the follow-up time
- Indicator e_i indicates the status at the end of the follow-up:

$$e_i = \begin{cases} 1 & \text{Patient } i \text{ experiences an event} \\ 0 & \text{Patient } i \text{ does not experiences an event; the event is censored} \end{cases}$$

5.3. Selected data characteristics and information

For clarity, to illustrate the data characteristics and information of the selected data, we use the selected index time as the first index time.

The resulting dataset consists of the selected patients, their individual follow-up time, and the events that occurred. In addition, patient demographics and characteristics, represented by the vectors $\mathbf{B}_i$ and $\mathbf{C}_i$ introduced in the previous section, are included. The covariates in this vector, such as age and various clinical measurements, were originally measured at multiple time points. However, the values used in this dataset are those measured at baseline.

Figure 5.4 shows the plots of the timeline of events by patient, split by the EU North-West and EU South-Central patients. For each patient, the line represents the follow-up time in years, starting from the start of their follow-up and ending at an event or when they were lost to follow-up. A blue line indicates a male patient and a pink line a female patient. If a patient experienced an event, it is indicated by a cross (death), a circle (liver transplantation) or a triangle (SBD). When no figure is present at the end of a line, it means the patient is lost to follow-up and has experienced no event. The results show, for example, that very few patients died.
Figure 5.5 presents the same data as Figure 5.4, but with the x-axis now representing patients' ages rather than survival time. Each line shows a patient's follow-up period, starting from their age at baseline. The figure reveals that most patients begin follow-up at a young age, though some start later, for example, after the age of 10.

Table 5.1 gives the baseline demographics and characteristics of the selected data. The baseline demographics of the PFIC2 patients are also given in the left part of Table 5.2. Continuous variables are expressed as median values with interquartile range (IQR) and were compared using the Mann-Whitney U test, for continuous non-normal outcomes, or Student's t-test, for continuous normal outcomes. Categorical variables are expressed as numbers and percentages and were compared using the Chi-square test. For most variables, the results show that there are imbalances present between the patients with PFIC2 from North-West European and South-Central European populations. Therefore, it is needed to perform some weighting methods to assess the balance of these covariates.

(a) EU North-West

(b) EU South-Central

Figure 5.4: The timeline of events by patient, split by the EU North-West and EU South-Central patients. A line represents the individual follow-up from the start until the end in years. The events, if observed, are indicated by a cross (death), liver transplantation (circle) and a triangle (SBD). The patients are grouped according the similar events. The pink lines indicate female patients, and the blue lines indicate male patients.
*Patient IDs were randomly assigned and are not traceable to the original IDs, ensuring anonymity.

(a) EU North-West

(b) EU South-Central

Figure 5.5: The timeline of events by patient, split by the EU North-West and EU South-Central patients, on the age scale. A line represents the years of follow-up, indicated by the age of the patient. The events are indicated by a cross (death), liver transplantation (circle) and a triangle (SBD). The patients are grouped according the similar events. The pink lines indicate female patients, and the blue lines indicate male patients.
*Patient IDs were randomly assigned and are not traceable to the original IDs, ensuring anonymity.

5.4. Balancing data cohorts using weighting methods

As explained in the previous chapter, some imbalances are present in the background characteristics. These imbalances can cause bias in the regional selection and, therefore, could give a biased result when comparing the two regions. The goal is therefore to assess the balance of the characteristics in the regional groups and obtain a better idea of the effect of the region on the event-free survival.

To adjust for the imbalances, the IPTW procedure was performed. The method behind this procedure is explained in Section 3.2. We start by calculating the propensity score using logistic regression as the probability of being in region North-West Europe versus region South-Central Europe. The known baseline confounders that are included in the model as covariates are the covariates in the vectors $\mathbf{B}_i$ and $\mathbf{C}_i$ in Section 5.2, but we will list them here again:

- Age at baseline
- Sex
- Clinical measurements of ALT
- Clinical measurements of sBA levels
- Clinical measurements of total bilirubin
- Type of PFIC*

For the liver biochemistry covariates, the logarithms of the baseline values were included in the model. *The variable for indication of the type of PFIC is not included in the model as a covariate, but the analysis needs to be stratified for the PFIC type, which means that the analysis is done separately for the patients with PFIC1 and PFIC2. The results for PFIC2 are shown first.

The propensity score for the i-th PFIC2 patient is defined as:

$$e(\mathbf{X}_i) = P(Z_i = 1|\mathbf{X}_i), \tag{5.1}$$

where

- $X_i = (\text{age}, \text{sex}, \log(\text{ALT}), \log(\text{sBA}), \log(\text{total bilirubin}))$
- and $Z_i = \begin{cases} 1 & \text{Region North-West Europe} \\ 0 & \text{Region South-Central Europe} \end{cases}$.

Since the balance is already present by only adding the main effects in the logistic regression model, we do not need to revisit the propensity model by including, e.g. interactions, transformations or splines. When the balance is not present by only adding the main effects of these confounders in the logistic regression model, the propensity model needs to be revisited by including, for example, interactions, transformations or splines [3] [5]. Some analysis has already been performed to see which interactions could then, for example, be added. The correlations between the baseline confounders, which are added to the model, have been checked to see if there are some variables with a high correlation in this dataset. If this is the case, the additional interactions of these variables can be added to the model, because then the confounding effect of one of these variables will vary by the other variable [31]. Figure 5.6 gives the correlation plot for the baseline confounders. Overall, the variables exhibit low to moderate correlations. However, a bit stronger correlations are observed among the biochemical parameters.

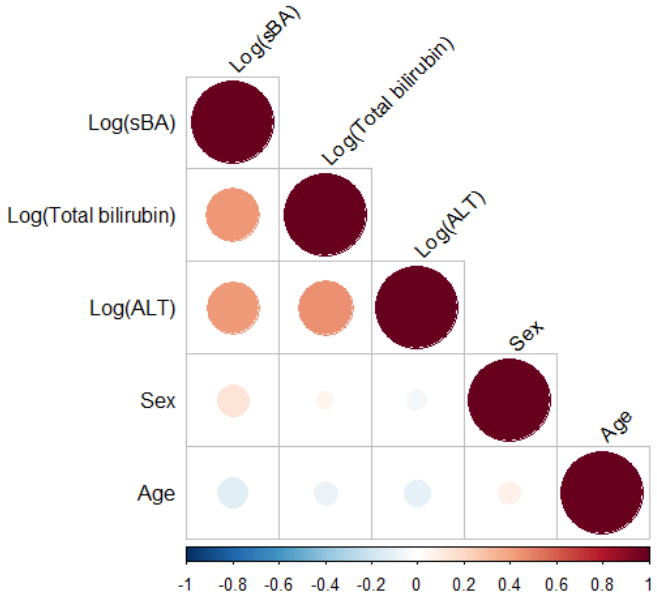


Figure 5.6: Correlation plot for the baseline confounders for PFIC2 patients

The logistic regression model gives the probability, or propensity score, of belonging to a certain region for each patient given their characteristics [10]. The distribution of the propensity scores by the two regional groups (for PFIC2 patients) is given in Figure 5.7. Evaluation of these distributions checks for sizeable overlap among the groups, demonstrating whether the groups are comparable. Figure 5.7 shows a large overlap in propensity scores.

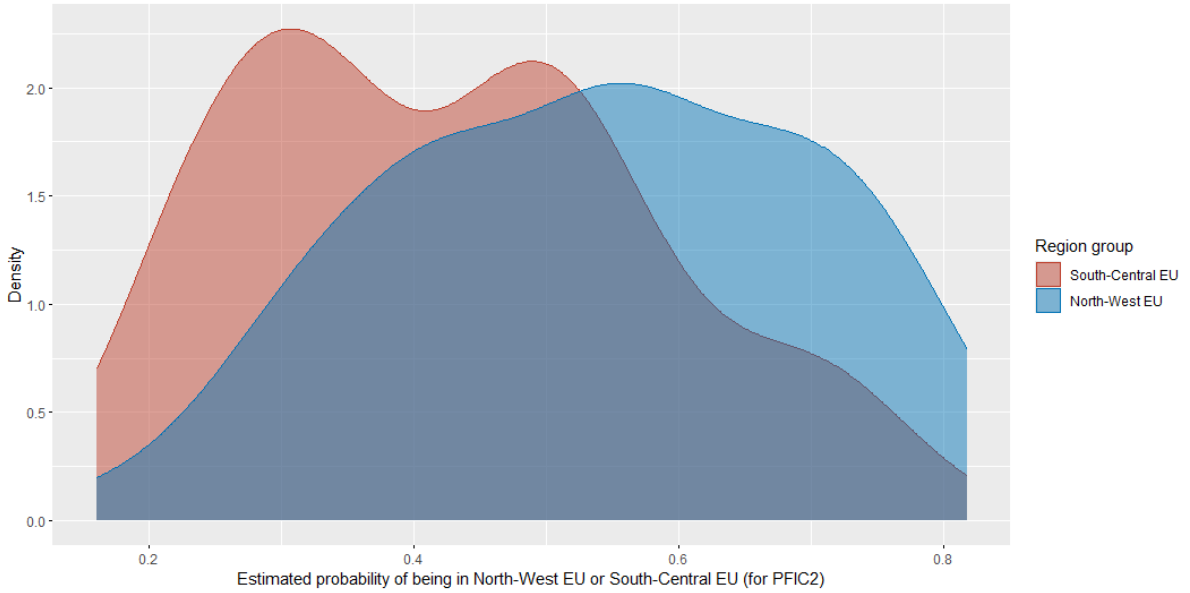


Figure 5.7: Distribution of the propensity scores by the groups North-West EU and South-Central EU, for PFIC2 patients using first visit index time.

From each calculated propensity score, the relative inverse probability of treatment weight and the stabilised inverse probability of treatment weight are calculated as described in equations 3.16 and 3.20. The first check of the weights is performed by plotting the boxplot of the stabilised weights,

shown in Figure 5.8. The weights are already quite small, so there is no need for weight truncation. The stabilised weights for each patient were used as the final IP weights.

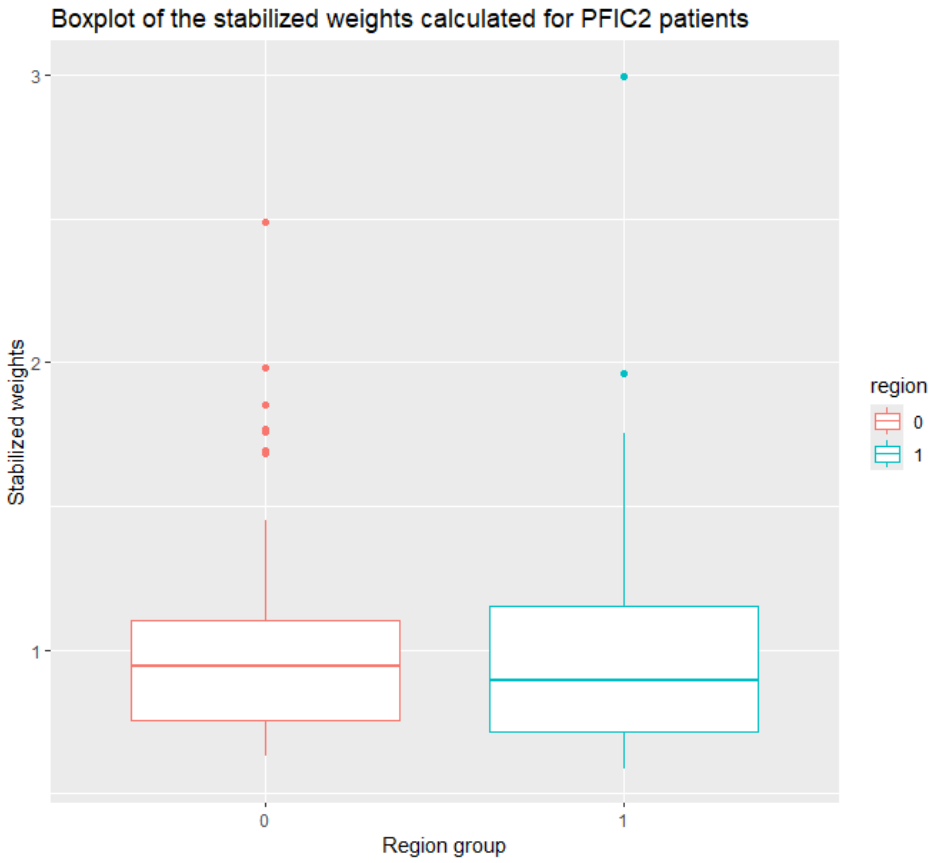


Figure 5.8: Boxplot of the stabilised weights for PFIC2 patients, split by the regional groups. Region 0 indicates the North-West region, and region 1 indicates the South-Central region.

After assigning each patient their corresponding IP weight, the covariate balance between the two groups is checked. This is done by assessing the standardised mean differences of the baseline characteristics included in the propensity score model before and after weighting. These standardised mean differences are displayed in Figure 5.9. The blue dots indicate the standardised mean differences between the North-West and South-Central European PFIC2 patients before weighting, and the red squares after weighting. The vertical black solid line represents the borders for the standardised mean difference of ± 0.10 and the dotted lines for ± 0.05. Balance is achieved when the standardised mean difference of all covariates is within the borders ± 0.10.

The results in Figure 5.9 show that before the weighting, balance was indeed not present for all covariates, except for the sBA covariate. This corresponds with our expectations based on the values in Table 5.2. The results after the IPTW adjustment show that the standardised mean difference of all covariates is within the borders ± 0.05, which indicates a good covariate balance. The Figures A.1 and A.2 in the Appendix A give figures in which balance in density and eCDF plots is displayed for the covariates added to the propensity score model, before and after IPTW adjustment. Perfectly overlapping lines indicate good balance. These figures also indicate that, after adjustment, the covariates are balanced.
Since the balance is already present by only adding the main effects in the logistic regression model, we don't need to revisit the propensity model by including, e.g. interactions, transformations or splines [3] [5].

In addition to the previously discussed baseline characteristics before weighting, Table 5.2 additionally presents the baseline characteristics after applying the IP weights. After IPTW adjustment, the sum

of weights in the North-West Europe region is 64.1, and in the South-Central Europe region is 68.9. The sample sizes of the two regions in the pseudo data differ only slightly from the original sample sizes. Which is correct since we have used the stabilised weights, explained in Section 3.2. For the comparison of the continuous and categorical variables after IPTW adjustment, the weighted Mann-Whitney U test, weighted Student's t-test and weighted Chi-square test are used. The p-values before and after IPTW adjustment indicate that, following adjustment, there is stronger evidence in favour of the null hypothesis, suggesting no significant difference in the specific variable between the two regional groups.

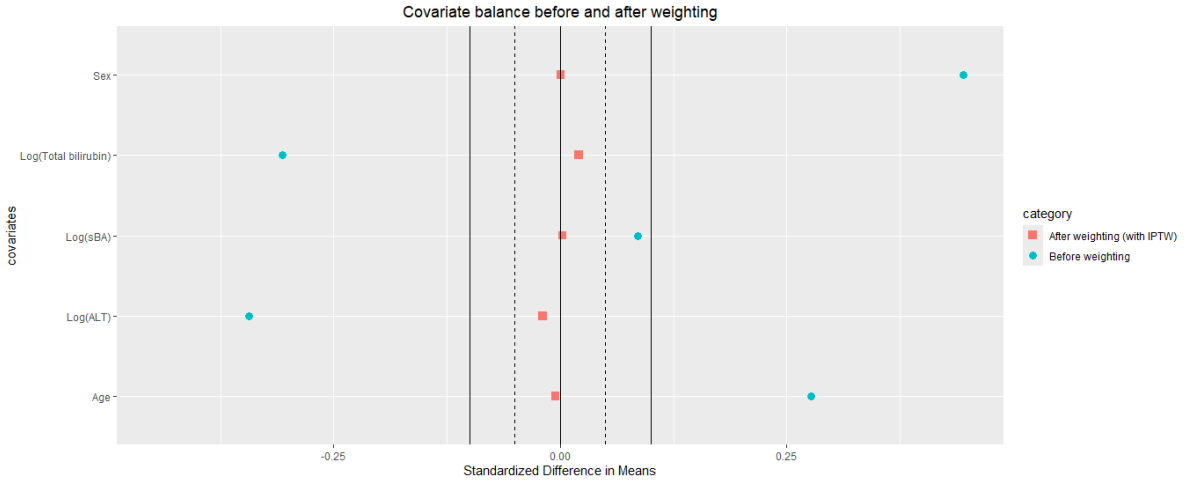


Figure 5.9: Plot of the covariance balance before and after weighting. The standardised mean differences are plotted for each covariate added to the propensity score model.

A similar analysis was conducted for PFIC1 patients. Extensive efforts were made to evaluate the balance within the PFIC1 group. However, the sample sizes in this group are limited, with only $n = 18$ for the North-West European region and $n = 14$ for the South-Central European region. Due to the small sample size and therefore insufficient statistical power, the balance could not be achieved. As a result, we have decided to exclude the PFIC1 patients from the final analysis in this thesis.

5.5. Results comparison of long-term outcomes

This section presents the results of the comparison of the event-free survival functions of North-West Europe versus South-Central Europe of PFIC2 patients. After presenting the results using the first visit index time, we also give the sensitivity analysis using the other possible index times.

The hypothesis for this comparison is defined as:

H_0 : The distribution of the event-free survival times is the same for North-West Europe and South-Central Europe:
$S_{EU_{SC}}(t) = S_{EU_{NW}}(t)$ for all $t \geq 0$

H_1 : The distribution of the event-free survival times is not the same for North-West Europe and South-Central Europe:
$S_{EU_{SC}}(t) \neq S_{EU_{NW}}(t)$ for at least one $t \geq 0$.

To perform this comparison, we first use the adjusted Kaplan-Meier estimator to compute the survival functions. The calculated IP weights derived in the previous section are used. The Kaplan-Meier curve after adjustment for age, sex, ALT, sBA and total bilirubin using IPTW is given in Figure 5.10. From the Kaplan-Meier curve, we can already see that the event-free survival functions of the two regions are very similar. To obtain a measure of the regional effect on the survival time distributions, we need to calculate the hazard ratio between the two European regions. Here, the weighted Cox regression model is used.

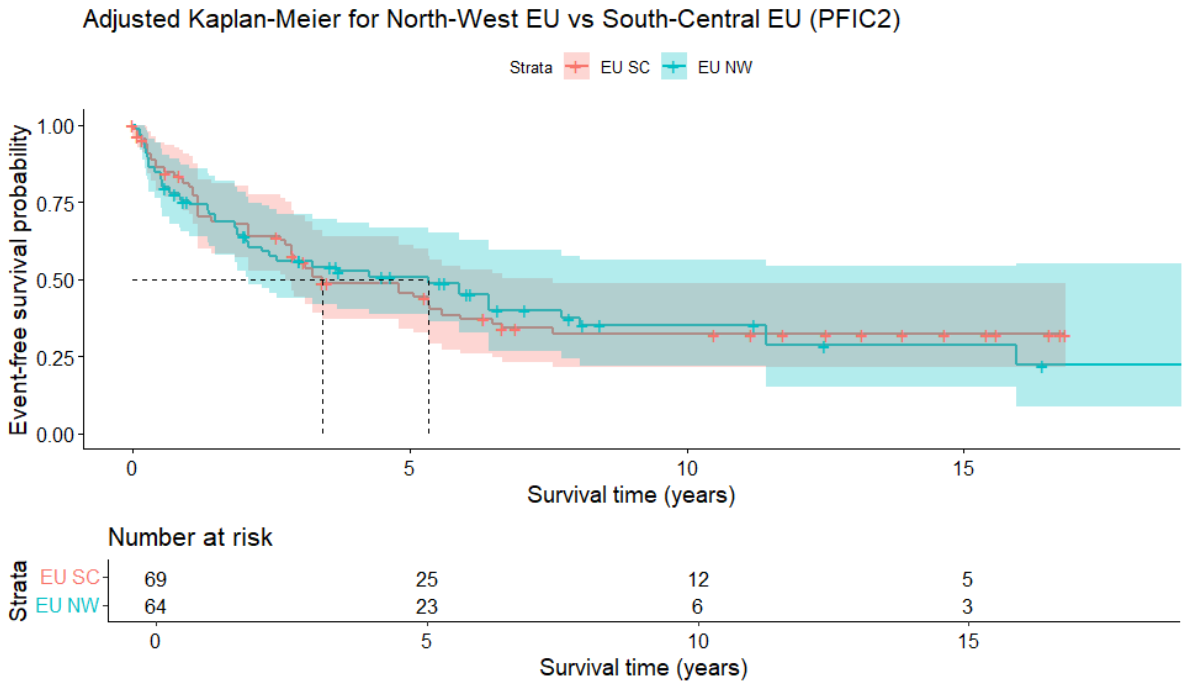


Figure 5.10: The adjusted Kaplan-Meier curve for the comparison of the survival time distribution for North-West Europe versus South-Central Europe for PFIC2 patients.

Before applying the weighted Cox regression model, we must verify the proportional hazards assumption. As explained in Section 3.1.4, the PH assumption requires that the hazard ratio is constant over time, so that the effect of a covariate on the risk of an event is constant over time. This is assessed using both statistical tests and graphical diagnostics based on the scaled Schoenfeld residuals [23]. Under the PH assumption, these residuals should be independent of time; therefore, any systematic pattern in their plot over time may indicate a violation.

The results of the test are given in Figure 5.11. The plot shows the Schoenfeld residuals (β) for the regional covariate. The smooth black solid curve is relatively flat, although with a slight rise at the end, but the wide confidence interval suggests this rise is not statistically meaningful. Therefore, the effect of the regional covariate appears to be constant over time. The p-value of the test is also not statistically significant, so we can conclude that there is no violation of the proportional hazards assumption.

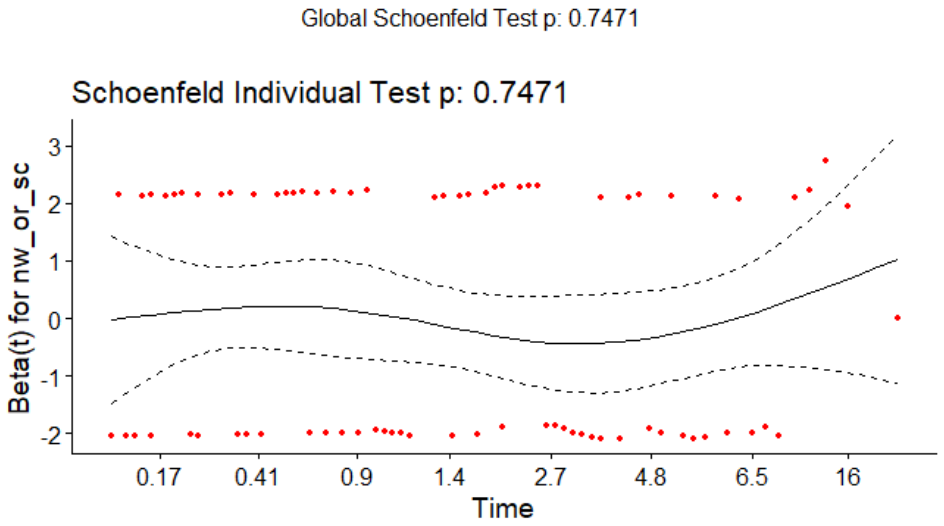


Figure 5.11: Schoenfeld residuals for the regional covariate.

The weighted Cox regression model gives the adjusted hazard ratio (HR) of 0.993, with a 95% confidence interval (CI) of 0.614 to 1.606, and the p-value of the weighted log-rank test of 0.971. The adjusted HR is very close to 1, which suggests almost no difference in the hazard between the two regions in Europe. In addition, the p-value of the weighted log-rank test is much higher than the significance threshold of 0.05, therefore, the result is not statistically significant, and we do not reject the null hypothesis of having the same survival times distribution between the European regions.

5.5.1. Sensitivity Analysis

As described in Section 5.1.2, there are different possibilities to determine the index time. To determine the degree to which the selection procedure affected the event-free survival time distribution, we will perform a sensitivity analysis for the different index times, selecting a random eligible visit per individual and selecting the last eligible time per individual. The results are given in Figure 5.12. Across all three indexing strategies (first, random, and last eligible visit), the results are not statistically significant, which suggests no statistical association between the regional effect and the event-free survival, regardless of how the index time is defined.

The sensitivity analysis also consists of the results of the unadjusted model, the model where there is no adjustment for some covariates. There is a small (negligible) difference compared to the adjusted IPTW model.

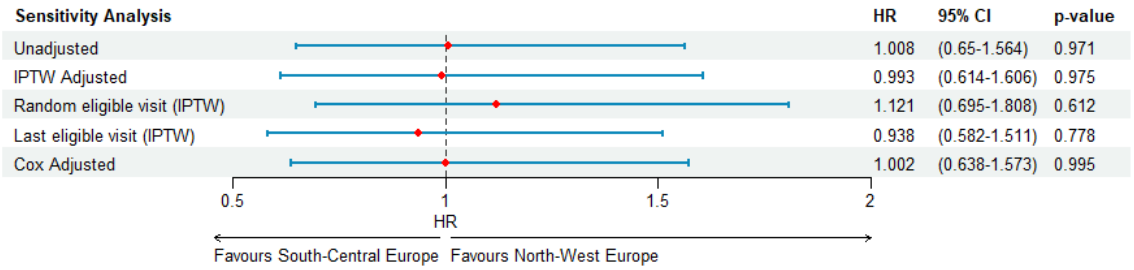


Figure 5.12: PFIC2 sensitivity analysis; for the unadjusted, IPTW and Cox adjusted models, the first visit index time is used. The HR gives the hazard for North-West Europe relative to the hazard of South-Central Europe.

It is also possible to compare the event-free survival function of the two European regions using a Cox proportional hazards model on its own. In this approach, the regional indicator is included as a binary

covariate, along with the same set of covariates originally used in the propensity score model. The results of this Cox model are also presented in Figure 5.12 under the label ”Cox Adjusted”. These findings are consistent with those obtained from the other analytical approaches.

5.6. Permutation test

This section gives the results of a different type of test, the permutation test, for testing the null hypothesis of no difference between regional groups. A permutation test is a non-parametric procedure for determining statistical significance based on rearrangements of the labels of a dataset [15]. A permutation test builds sampling distributions by resampling the observed data, e.g. by assigning each individual to a different group.

The null hypothesis for the permutation test is: H_0 : the group labels assigning samples to classes are interchangeable [15].
The original test statistic, for the original group division, is compared with the sampling distribution of permutation values. These permutation values are computed similarly to the test statistic, however, under a random rearrangement (permutation) of the group labels in the dataset [24]. The significance of a permutation test is represented by its p-value, which is the probability of obtaining a result at least as extreme as the test statistic, given that the null hypothesis is true [24]. Obtaining a significant p-value indicates that the group labels are not interchangeable and therefore that the original group labels are relevant to the data. The p-values are calculated by performing a number N of permutations and computing the proportion of the N permutation values that are at least as extreme as the original test statistic obtained from the data before the permutations [24].

We have performed the permutation test for the IPTW weighted model. This means that, after every permutation, the IPTW weights are recalculated. These IPTW weights are then used to create a distribution of weighted log-rank statistics under the null hypothesis:
H_0 : The group labels for assigning patients to the region of North-West Europe or South-Central Europe are interchangeable. Which, in other words, is the hypothesis that there is no difference in event-free survival between North-West Europe and South-Central Europe.

We also performed the permutation tests for the unadjusted model, and the IPTW adjusted and unadjusted model with the random visit index time as a sensitivity analysis. All permutation tests have been performed with 1000 permutations, and the results are given in Table 5.3. Figures B.1 and B.2 in the Appendix B show the histogram of the unweighted and weighted model with the first visit index time and random visit index time, respectively. The red dotted line indicates the original log-rank statistic, and the histogram bars give the sample distribution of the log-rank test statistics after permutations. It is visible that the red line falls within the sample distribution for all models.

Table 5.3: Results of the permutation tests

	Original Log-Rank Statistic	Permutation p-value
Unadjusted (first eligible visit)	0.00137	0.98
IPTW adjusted (first eligible visit)	0.00095	0.981
Unadjusted (random eligible visit)	0.0933	0.745
IPTW adjusted (random eligible visit)	0.269	0.574

From the p-values of the permutation tests, we can also conclude that for all models, the p-values are not statistically significant and therefore indicate strong evidence for the null hypothesis, that the group labels for assigning patients to the region of North-West Europe or South-Central Europe are interchangeable. These results match the results from the weighted Cox regression model in Section 5.5.

Table 5.1: Baseline demographics and characteristics of selected data using the first index time.

		EU North-West (n = 82)		EU South-Central (n = 83)	
		PFIC1	**PFIC2**	**PFIC1**	**PFIC2**
Number of patients		18 (22.0)	64 (78.0)	14 (16.9)	69 (83.1)
Sex	Male	12 (66.7)	26 (40.6)	8 (57.1)	43 (62.3)
	Female	6 (33.3)	38 (59.4)	6 (42.9)	26 (37.7)
Age at baseline, in years		2.4 (1.6–6.3)	2.3 (1.3–7.3)	1.8 (1.7–2.9)	1.9 (1.3–4.2)
Total Bilirubin at baseline, in $\log_{10} \times ULN$		0.3 (-0.2–0.8)	0.2 (-0.03–0.7)	0.6 (0.4–0.8)	0.5 (0.04–0.9)
sBA at baseline, in $\log_{10} \times ULN$		1.4 (1.3–1.5)	1.5 (1.3–1.6)	1.4 (1.2–1.5)	1.5 (1.3–1.6)
ALT at baseline, in $\log_{10} \times ULN$		0.01 (-0.1–0.2)	0.3 (0.1–0.5)	0.2 (0.1–0.3)	0.4 (0.1–0.7)
Number of patients with first event	SBD	6 (33.3)	22 (34.4)	6 (42.9)	20 (29.0)
	LTx	2 (11.1)	15 (23.4)	4 (28.6)	21 (30.4)
	Death	0 (0)	2 (3.13)	0 (0)	1 (1.45)
Survival time, in years		3.0 (0.4–9.7)	2.4 (0.7–5.9)	1.7 (0.4–3.9)	2.9 (0.8–6.3)

Table 5.2: Baseline characteristics of PFIC2 patients before and after IPTW adjustment

Variables		Before IPTW adjustment			After IPTW adjustment		
		EU North-West (n = 64)	EU South-Central (n = 69)	P-value	EU North-West (n = 64) Sum of weights = 64.1	EU South-Central (n = 69) Sum of weights = 68.9	P-value
Sex	Male	26 (40.6)	43 (62.3)	0.012	33.5 (52.3)	36.0 (52.2)	0.998
	Female	38 (59.4)	26 (37.7)		30.6 (47.7)	32.9 (47.8)	
Age at baseline, in years		2.3 (1.3-7.3)	1.9 (1.3-4.2)	0.256	2.0 (1.3-5.0)	2.6 (1.3-5.1)	0.670
Total bilirubin at baseline, in $\log_{10} \times ULN$		0.2 (-0.03-0.7)	0.5 (0.04-0.9)	0.082	0.4 (0.04-0.8)	0.3 (-0.1-0.9)	0.907
sBA at baseline, in $\log_{10} \times ULN$		1.5 (1.3-1.6)	1.5 (1.3-1.6)	0.6227	1.5 (1.3-1.6)	1.5 (1.3-1.6)	0.987
ALT at baseline, in $\log_{10} \times ULN$		0.3 (0.1-0.5)	0.4 (0.1-0.7)	0.052	0.4 (0.2-0.6)	0.3 (0.1-0.6)	0.912

6

Conclusion and Discussion

In this chapter, we present the main conclusions of our research in Section 6.1, Section 6.2 reflects on our findings and discusses the key challenges encountered. And finally, suggestions for future work are given.

6.1. Key findings

6.1.1. Longitudinal trajectories of the biochemical parameters sBA levels, and platelet count

To address the first objective of this thesis, to determine and identify similarities of trajectories of the relevant biochemical parameters, sBA levels and platelet counts in patients with PFIC2, the latent class linear mixed model was used. Two LCLMMs are used to identify the subgroups of sBA levels and platelet counts separately. The models both identify two classes for the sBA model and two classes for the PLT model:

Results for the sBA model:
Class sBA1 shows a slight increase in sBA levels following an initial decline. In contrast, Class sBA2 exhibits consistently low levels from an early age onward. The majority of the patients, 84%, are classified in Class sBA1, while the remaining 16% belong to Class sBA2.
From a clinical perspective, Class sBA2 is an unexpected finding in patients with PFIC2, as they typically exhibit elevated sBA levels. Detailed analysis of the patients of this group reveals a group of patients with the episodic form of PFIC.

Results for the PLT model:
Class PLT1 shows a slow reduction in platelet count values, whereas Class PLT2 exhibits a more rapid decrease. Again, the majority of the patients, 87%, are classified in Class PLT1, and the remaining 13% belong to Class PLT2.
From a clinical perspective, the patients in Class PLT2 could indicate patients with a more rapid development of an overactive spleen.

By combining the findings from both latent class mixed models, we conclude that this approach effectively identifies longitudinal patterns of sBA levels and platelet counts in patients with PFIC2. These patterns highlight significant heterogeneity in the progression of laboratory parameters over time. These insights enhance the understanding of the disease and hold potential for improving patient care. Specifically, they may support the development of targeted therapeutic strategies, enabling earlier and more effective intervention for patients at higher risk.

6.1.2. Comparison of event-free survival in PFIC patients of cohorts within Europe

The second objective of this thesis was to perform a comparison of event-free survival between two regional cohorts of PFIC patients within Europe. This comparison can indicate the statistical regional

effect on event-free survival. To assess this second objective, a weighted survival analysis has been performed, using the combination of the IPTW and the Kaplan-Meier estimator and the Cox regression model.

The results from the weighted Log-rank test and weighted Cox regression model indicate that the null hypothesis of having the same distribution of the event-free survival times for North-West Europe and South-Central Europe is not rejected. The additional permutation test gives the same results. This indicates that there is no statistical regional effect on the event-free survival of PFIC2 patients, at least for the regions North-West Europe versus South-Central Europe.

Sensitivity analyses have been performed to check if different options for choosing the index time will change the results. The results of the sensitivity analyses of the weighted Cox regression model and the permutation test both indicate that for all chosen index times, the p-values for the null hypotheses are still not statistically significant. There are some slight differences, which will be discussed in the next section.

6.2. Discussion

In this section, we reflect on some key findings and discuss aspects of the study that could be improved.

The results of the sensitivity analysis in Section 5.5.1 indicate a minimal change in the hazard ratio and p-value after IPTW adjustment compared to the unadjusted model. The result of the permutation test in Table 5.3 also indicates this minimal difference in the permutation p-values. This minimal change suggests that the influence of measured confounding on the regional effect on the event-free survival may be limited. Alternatively, the small difference could also indicate that the IPTW procedure was not sufficiently effective in correcting for imbalance, perhaps due to some unobserved confounders. If unobserved confounders are present, residual bias can remain no matter which method is used [43]. So it could be the case that some confounders are missed and not added to the propensity score model.

The confidence intervals of the hazard ratios in the sensitivity analysis are very wide. This reflects uncertainty and suggests potentially low statistical power. As a result, the possibility of a moderate statistical regional effect cannot be entirely excluded. However, the consistency of the findings across all index time definitions, each having hazard ratios close to 1 and lacking statistical significance, strengthens the case for the conclusion that there is no meaningful statistical regional effect on the event-free survival in this dataset.

A complication we ran into during my thesis was the interpretation of the data. Given that the dataset contains data of patients of a rare disease, it involves considerable medical terminology and requires specific medical knowledge to fully understand. Since the data are collected by different centres, there could be inconsistencies in how some data is filled in. This gave mostly some difficulties in understanding the units of the biochemical parameters and to make sure these are all consistent to use this data in the analysis. This process might have been streamlined if the data had been pre-checked by some laboratory technicians to ensure consistency in how each value was recorded and labelled.

6.3. Suggestions for future work

Based on the challenges and limitations encountered in this research, we propose two directions for future research.

The original aim of the comparison of cohorts in terms of event-free survival was to compare the dataset used in this thesis with a dataset of treated patients. This comparison could have provided insights into the impact of treatment on event-free survival of PFIC patients. However, due to the unavailability of the additional dataset within the project timeline, the focus was shifted to comparing two regional cohorts within Europe to allow the use of the same methodology. Even though the original aim could not be achieved, the analytical methods investigated in this project remain applicable when the necessary dataset of treated patients is available. As soon as this is available, this approach, together with the corresponding code, will be ready for immediate use. Therefore, the primary recommendation for future research is to apply this methodology to the originally intended comparison with the dataset of treated patients as soon as it is available.

Another limitation lies in the variable selection process for the propensity score model used in IPTW. When acceptable covariate balance is not achieved after only including the main effects, the propensity score model must be revisited, often by adding interaction terms. According to the literature, in practice, this process relies heavily on clinical knowledge and iterative trial-and-error, evaluating whether added interactions or nonlinear transformations improve balance, and adjusting the model accordingly. We believe this approach can be made more efficient and less reliant on the clinicians.
Several variable selection techniques have been proposed to guide the process of variable selection in the propensity score model, such as the LASSO method [44]. However, fitting a propensity score model using LASSO regularisation with a shrinkage parameter selected via cross-validation typically prioritises prediction accuracy of treatment assignment, rather than optimising covariate balance or treatment effect estimation [6], which are the primary objectives of using the propensity score model in this thesis. Nonetheless, recent literature shows that LASSO regularisation with a shrinkage parameter selected to directly target covariate balance is feasible [6]. We recommend that future research explore this approach to improve variable selection for propensity score modelling in IPTW.

References

[1] Victoria Allan et al. “Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants”. In: *Journal of comparative effectiveness research* 9.9 (2020), pp. 603–614.

[2] Carolyn J Anderson and Leslie Rutkowski. “Multinomial logistic regression”. In: *Best practices in quantitative methods*. SAGE Publishing, 2008, pp. 390–409.

[3] Peter C Austin. “An introduction to propensity score methods for reducing the effects of confounding in observational studies”. In: *Multivariate behavioral research* 46.3 (2011), pp. 399–424.

[4] Peter C Austin. “Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples”. In: *Statistics in medicine* 28.25 (2009), pp. 3083–3107.

[5] Peter C Austin and Elizabeth A Stuart. “Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies”. In: *Statistics in medicine* 34.28 (2015), pp. 3661–3679.

[6] Guilherme WF Barros, Marie Eriksson, and Jenny Häggström. “Performance of modeling and balancing approach methods when using weights to estimate treatment effects in observational time-to-event settings”. In: *Plos one* 18.12 (2023), e0289316.

[7] David A Binder. “Fitting Cox’s proportional hazards models from survey data”. In: *Biometrika* 79.1 (1992), pp. 139–147.

[8] Ashley L Buchanan et al. “Worth the weight: using inverse probability weighted Cox models in AIDS research”. In: *AIDS research and human retroviruses* 30.12 (2014), pp. 1170–1177.

[9] Centraal Bureau voor de Statistiek (CBS). *Bevolkingsontwikkeling; regio per maand*. StatLine. 2024. URL: `https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37230ned/table?ts=1559810405300`.

[10] Nicholas C Chesnaye et al. “An introduction to inverse probability of treatment weighting in observational research”. In: *Clinical kidney journal* 15.1 (2022), pp. 14–20.

[11] Stephen R Cole and Miguel A Hernán. “Constructing inverse probability weights for marginal structural models”. In: *American journal of epidemiology* 168.6 (2008), pp. 656–664.

[12] Anne Davit-Spraul et al. “Progressive familial intrahepatic cholestasis”. In: *Orphanet journal of rare diseases* 4 (2009), pp. 1–12.

[13] Michael C Donohue et al. “Natural cubic splines for the analysis of Alzheimer’s clinical trials”. In: *Pharmaceutical statistics* 22.3 (2023), pp. 508–519.

[14] Anand Dutta et al. “Variability in the upper limit of normal for serum alanine aminotransferase levels: a statewide study”. In: *Hepatology* 50.6 (2009), pp. 1957–1962.

[15] Eugene Edgington and Patrick Onghena. *Randomization tests*. Chapman and Hall/CRC, 2007.

[16] Sara Enck. “Latent Class Linear Mixed Models: A General Approach Implemented via SAS Macro with a Tutorial for Clinical Researchers”. Ph.D. dissertation. Chapel Hill, NC: University of North Carolina at Chapel Hill, 2009. URL: `https://doi.org/10.17615/gqp1-yc24`.

[17] Osvaldo Espin-Garcia, Lizbeth Naranjo, and Ruth Fuentes-García. “A latent class linear mixed model for monotonic continuous processes measured with error”. In: *Statistical Methods in Medical Research* 33.3 (2024), pp. 449–464.

[18] Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Vol. 608. Springer, 2001.

[19] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[20] Miguel A Hernán and James M Robins. “Using big data to emulate a target trial when a randomized trial is not available”. In: *American journal of epidemiology* 183.8 (2016), pp. 758–764.

[21] KJ Jager et al. “Confounding: what it is and how to deal with it”. In: *Kidney international* 73.3 (2008), pp. 256–260.

[22] Geurt Jongbloed. *Modeling and analysis of time-to-event data*. Lecture notes, Course WI4220, Fall 2015, Delft University of Technology. 2015.

[23] David G Kleinbaum and Mitchel Klein. *Survival analysis a self-learning text*. Springer, 1996.

[24] Theo A Knijnenburg et al. “Fewer permutations, more accurate P-values”. In: *Bioinformatics* 25.12 (2009), pp. i161–i168.

[25] Shadi Kolahdoozan et al. “Upper normal limits of serum alanine aminotransferase in healthy population: a systematic review”. In: *Middle East journal of digestive diseases* 12.3 (2020), p. 194.

[26] Ilari Kuitunen et al. “Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review”. In: *BMC musculoskeletal disorders* 22.1 (2021), p. 489.

[27] Nan M Laird and James H Ware. “Random-effects models for longitudinal data”. In: *Biometrics* (1982), pp. 963–974.

[28] Katrin Madjar and Jörg Rahnenführer. “Weighted cox regression for the prediction of heterogeneous patient subgroups”. In: *BMC Medical Informatics and Decision Making* 21 (2021), pp. 1–15.

[29] Ashley Mehl et al. “Liver transplantation and the management of progressive familial intrahepatic cholestasis in children”. In: *World Journal of Transplantation* 6.2 (2016), p. 278.

[30] Ronald PJ Oude Elferink et al. “The molecular mechanism of cholestatic pruritus”. In: *Digestive Diseases* 29.1 (2011), pp. 66–71.

[31] C Fiorella Murillo Perez et al. “Greater transplant-free survival in patients receiving obeticholic acid for primary biliary cholangitis in a clinical trial setting compared to real-world external controls”. In: *Gastroenterology* 163.6 (2022), pp. 1630–1642.

[32] Mohamad Amin Pourhoseingholi, Ahmad Reza Baghestani, and Mohsen Vahedi. “How to control confounding effects by statistical analysis”. In: *Gastroenterology and hepatology from bed to bench* 5.2 (2012), p. 79.

[33] Cécile Proust and Hélène Jacqmin-Gadda. “Estimation of linear mixed models with a mixture of distribution for the random effects”. In: *Computer methods and programs in biomedicine* 78.2 (2005), pp. 165–173.

[34] Cécile Proust-Lima, Viviane Philipps, and Benoit Liquet. “Estimation of extended mixed models using latent classes and latent processes: the R package lcmm”. In: *Journal of statistical software* 78 (2017), pp. 1–56.

[35] Jason T Rich et al. “A practical guide to understanding Kaplan-Meier curves”. In: *Otolaryngology—Head and Neck Surgery* 143.3 (2010), pp. 331–336.

[36] Dimitris Rizopoulos. *Biostatistics II: Survival Analysis*. `https://www.drizopoulos.com/courses/EMC/EP03.pdf`. Lecture slides, Department of Biostatistics, Erasmus University Medical Center.

[37] Dimitris Rizopoulos. *Statistical Analysis of Repeated Measurements Data*. `https://www.drizopoulos.com/courses/EMC/CE08.pdf`. Lecture slides, Department of Biostatistics, Erasmus University Medical Center.

[38] Paul R Rosenbaum and Donald B Rubin. “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 70.1 (1983), pp. 41–55.

[39] Nagoud Schukfeh et al. “Normalization of serum bile acids after partial external biliary diversion indicates an excellent long-term outcome in children with progressive familial intrahepatic cholestasis”. In: *Journal of pediatric surgery* 47.3 (2012), pp. 501–505.

[40] Anshu Srivastava. “Progressive familial intrahepatic cholestasis”. In: *Journal of clinical and experimental hepatology* 4.1 (2014), pp. 25–36.

[41] Elizabeth A Stuart, Brian K Lee, and Finbarr P Leacy. “Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research”. In: *Journal of clinical epidemiology* 66.8 (2013), S84–S90.

[42] Terry M. Therneau. *survfit function — survival package documentation*. URL: `https://www.rdocumentation.org/packages/survival/versions/2.11-4/topics/survfit`.

[43] Felix Thoemmes and Anthony D Ong. “A primer on inverse probability of treatment weighting and marginal structural models”. In: *Emerging Adulthood* 4.1 (2016), pp. 40–59.

[44] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[45] Geert Verbeke and Emmanuel Lesaffre. “A linear mixed-effects model with heterogeneity in the random-effects population”. In: *Journal of the American Statistical Association* 91.433 (1996), pp. 217–221.

[46] Daan BE van Wessel et al. “Defining the natural history of rare genetic liver diseases: lessons learned from the NAPPED initiative”. In: *European journal of medical genetics* 64.7 (2021), p. 104245.

[47] Daan BE van Wessel et al. “Genotype correlates with the natural history of severe bile salt export pump deficiency”. In: *Journal of Hepatology* 73.1 (2020), pp. 84–93.

[48] Jun Xie and Chaofeng Liu. “Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data”. In: *Statistics in medicine* 24.20 (2005), pp. 3089–3110.

[49] Stanley Xu et al. “Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals”. In: *Value in Health* 13.2 (2010), pp. 273–277.

A

Results IPTW

This section presents additional results of the IPTW. Figure A.1 shows the density plots of the distributional balance for all the covariates added to the propensity score model. Figure A.2 shows the empirical CDF plots of the distributional balance for the four continuous confounding variables. In both figures, perfectly overlapping lines indicate good balance.

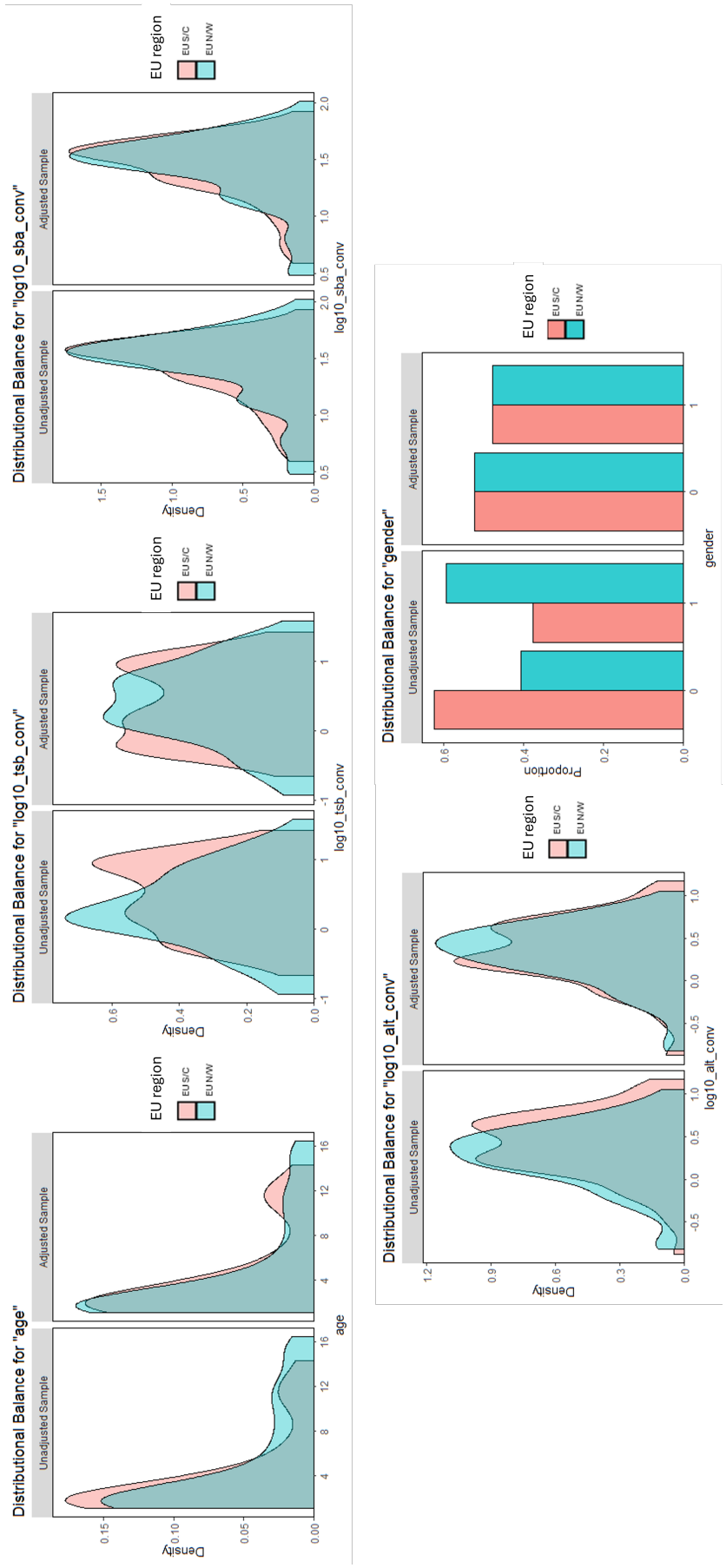


Figure A.1: Plots of the distributional balance for all the covariates added to the propensity score model separately. The density plots for the two European regions on the given covariate are given before and after IPTW adjustment. Perfectly overlapping lines indicate good balance.

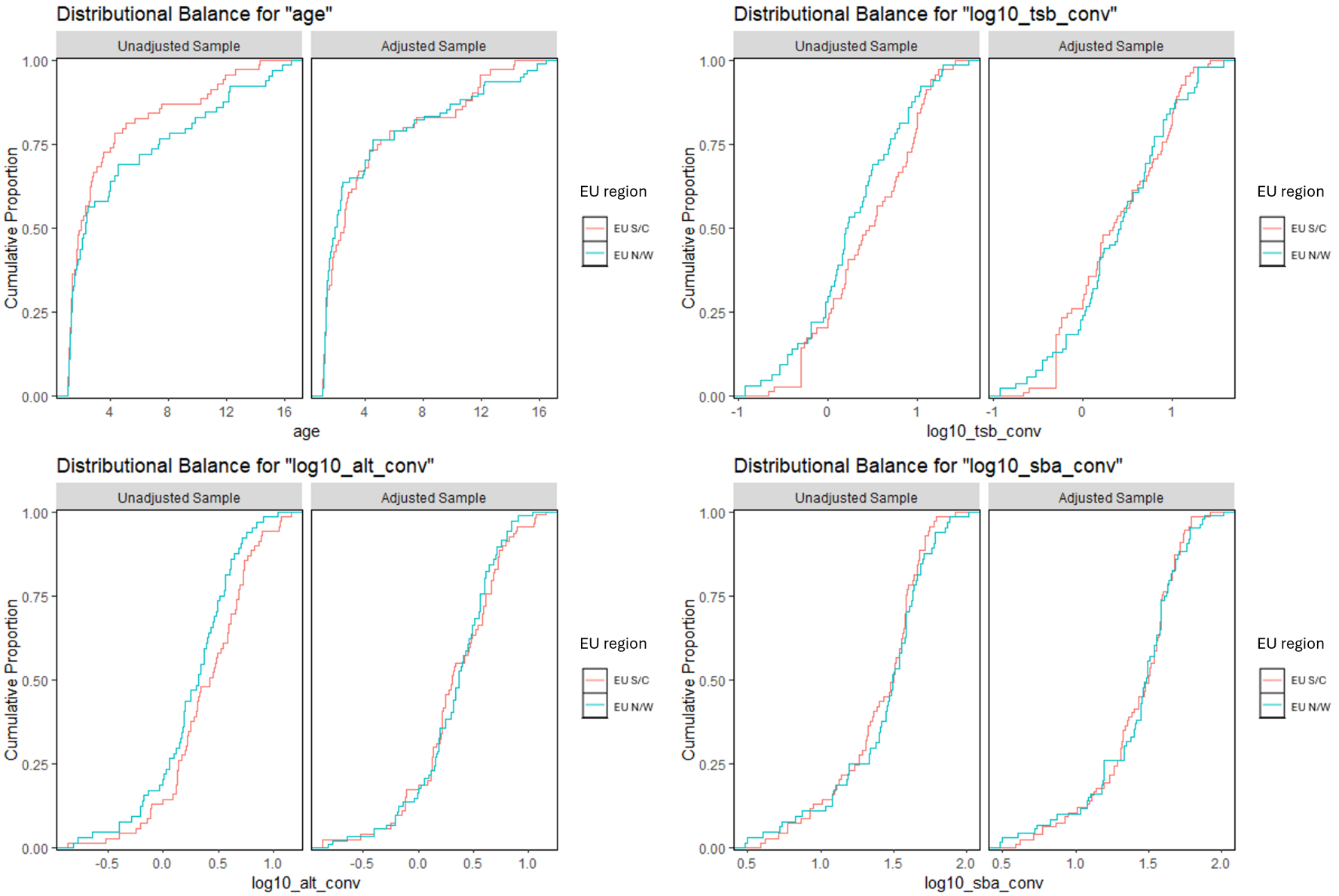
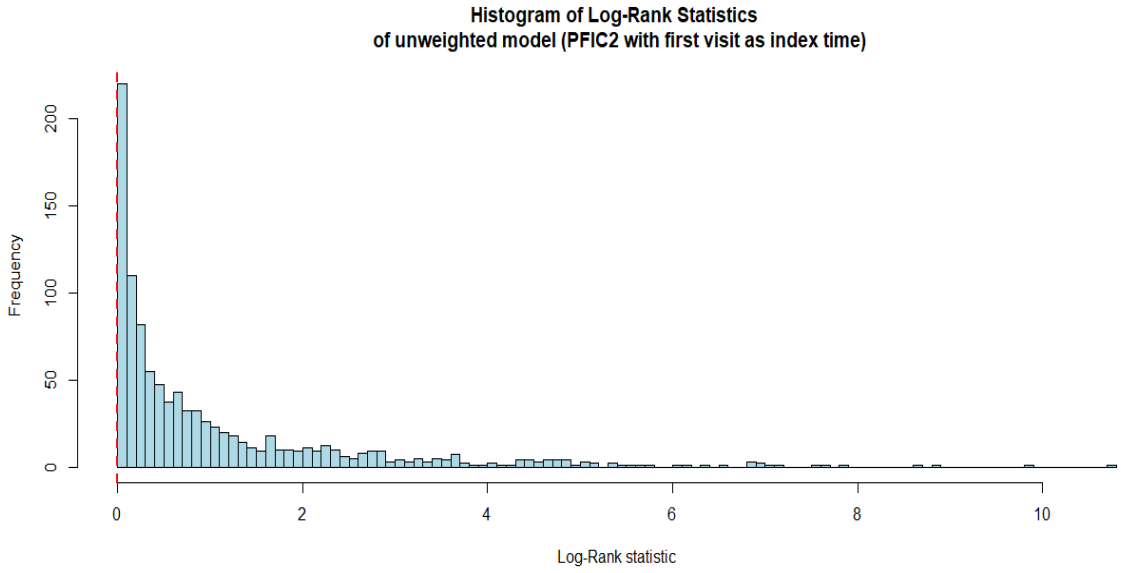


Figure A.2: Plots of the distributional balance for the 4 continuous covariates added to the propensity score model separately. The empirical CDF plots for the two European regions on the given covariate are given before and after IPTW adjustment. Perfectly overlapping lines indicate good balance.

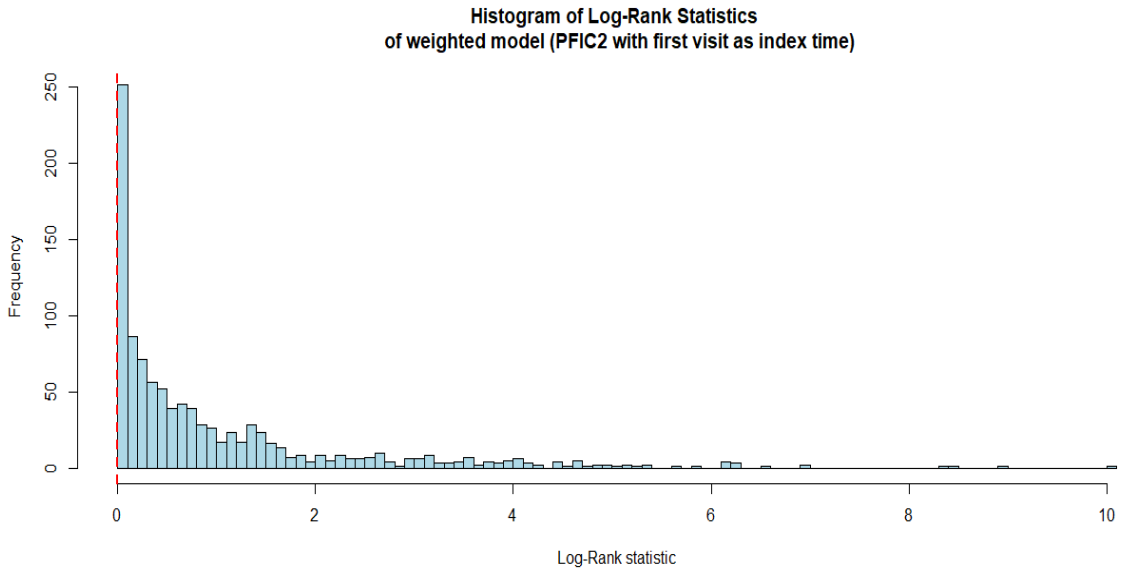
B

Results permutation tests

This section presents the results of the permutation test. The histograms of Log-Rank statistics of the unweighted and weighted models with the first visit index time and random visit index time are given in Figures B.1 and B.2 respectively. The red dotted lines indicate the original log-rank statistic, and the histogram bars give the sample distribution of the log-rank test statistic after 1000 permutations.

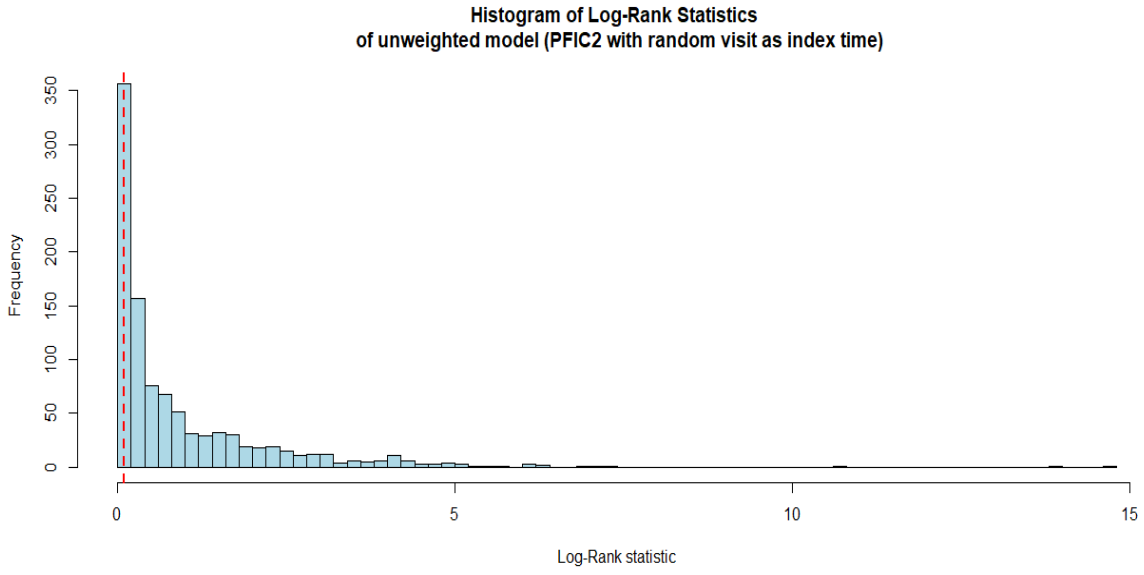


(a) Unweighted model

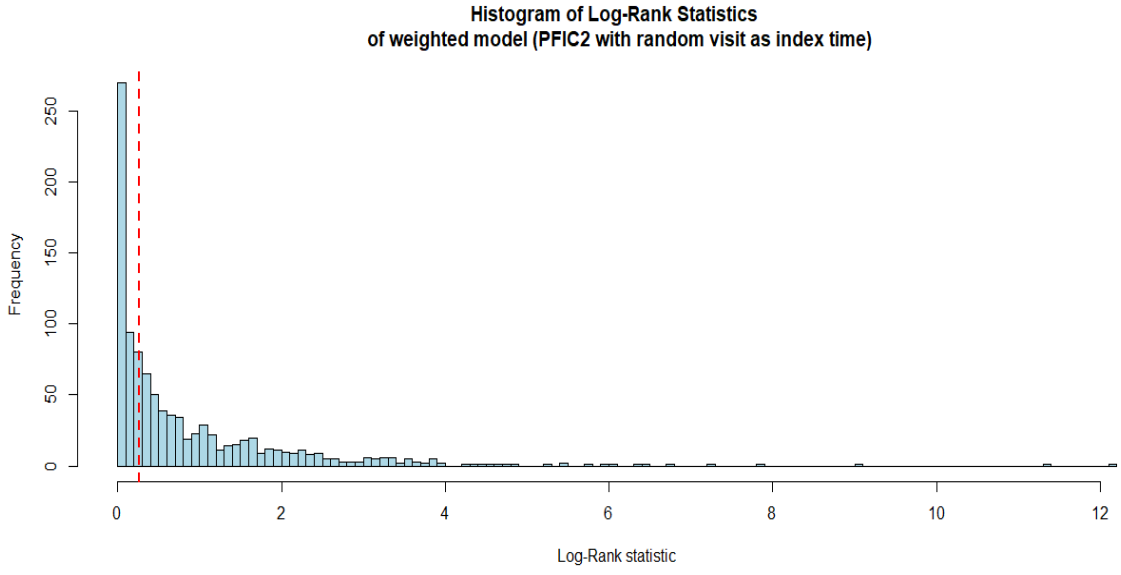


(b) Weighted model

Figure B.1: Histograms of Log-Rank statistics of the unweighted and weighted model with the first visit index time. The red dotted line indicates the original log-rank statistic, and the histogram bars give the sample distribution of the log-rank test statistics after permutations.



(a) Unweighted model



(b) Weighted model

Figure B.2: Histograms of Log-Rank statistics of the unweighted and weighted model with the random visit index time. The red dotted line indicates the original log-rank statistic, and the histogram bars give the sample distribution of the log-rank test statistics after permutations.

C

Poster Conference EASL

In this final section, Figure C.1 shows the presented poster at the European Association for the Study of the Liver (EASL) Congress 2025 in Amsterdam in May 2025.

Longitudinal Patterns of sBA Levels and PLT Counts in Patients with PFIC2 in the Global NAPPED Registry

EASL CONGRESS
7-10 May 2025
Amsterdam, the Netherlands

P Huisman[1,55], H.J. Verkade[2], A.D.E. de Groot[2], M. Nomden[2], A. Felzen[2], E. Gonzales[3], I. Jankowska[4], B. Shneider[5], E. Sokal[6], R.J. Thompson[7], J. Wang[8], L. Li[8], E. Jacquemin[3], P. Lipinski[4], P. Czubkowski[4], T. Grammatikopoulos[10], A. Kadaristiana[10], H. Arnell[11], B. Fischler[12], W. van der Woerd[13], E. Nicastro[14], L. d'Antiga[14], S. Alam[15], M. Shagrani[16,17], D. Broering[18], B.M. Kamath[19,20], A. Zellos[21], G. Gupte[22], P.L. Calvo[23], E. Grabhorn[24], D. Debray[25], M. Cananzi[26], M. Ruiz[27], C. Jimenez-Rivera[28], K. Kim[29], L. Hierro[30], G.M. Bartolo[30], Ö. Durmaz[31], G.F. Vogel[32,33], E. Sturm[34], G. Indolfi[35], A. Fabre[36,37], E. Shteyer[38], M. Hørby Jørgensen[39], J.S. Ko[40], N. Rock[41], A. Azaz[42], Y. Mozer-Glassberg[43], S. Ferreira[44], J. Breceli[45,46], G. Nebbia[47], I. Roilidis[48], A. Dezsofi-Gottl[49], E. Barkaoui[50], C. Tzivinikos[51], G. Jongbloed[55], P. Mian[52], B.E. Hansen[1,53,54], P. Miranda Afonso[1]

[1]Erasmus University Medical Center, NL; [2]University Medical Center Groningen; NL; [3]Biotère Hospital, FR; [4]Children's Memorial Health Institute, PL; [5]Baylor College of Medicine, US; [6]Cliniques St Luc, BE; [7]King's College London, UK; [8]Children's hospital of Fudan university, CN; [9]The Children's Memorial Health Institute, PL; [10]King's College Hospital, UK; [11]Karolinska Institutet, SE; [12]CLINTEC, Karolinska Institutet, SE; [13]Wilhelmina Children's Hospital, NL; [14]ASST Papa Giovanni XXIII, IT; [15]Institute of Liver and Biliary Sciences, IN; [16]King Faisal Specialist Hospital & Research Center, SA; [17]Alfaisal University, SA; [18]King Faisal Specialist Hospital & Research Center, SA; [19]Baylor College of Medicine, US; [20]The Hospital for Sick Children and the University of Toronto, CA; [21]National and Kapodistrian University of Athens, GR; [22]Birmingham Women's and Children's Hospital, UK; [23]Regina Margherita Children's Hospital, IT; [24]University Medical Center Hamburg Eppendorf, DE; [25]APHP-Necker Enfants Malades University Hospital, FR; [26]University Hospital of Padova, IT; [27]Hospices Civils de Lyon, FR; [28]Children's Hospital of Eastern Ontario, CA; [29]Asan Medical Center Children's Hospital, KR; [30]La Paz University Hospital, ES; [31]Istanbul Faculty of Medicine, TR; [32]Medical University of Innsbruck, AT; [33]Medical University of Innsbruck, AT; [34]University Children's Hospital Tuebingen, DE; [35]Meyer Children's, IT; [36]Aix Marseille University, FR; [37]Timone Enfant, FR; [38]Shaare Zedek Medical Center, IL; [39]Rigshospitalet Copenhagen University Hospital, DK; [40]Seoul National University College of Medicine, KR; [41]University Hospital of Geneva, CH; [42]Sheikh Khalifa Medical City, AE; [43]Schneider Children's Medical Center of Israel, IL; [44]Centro Hospitalar e Universitário de Coimbra, PT; [45]University of Ljubljana, SI; [46]University of Ljubljana, SI; [47]Fondazione Irccs Ca' Granda Ospedale Maggiore Policlinico, IT; [48]Aristotle University of Thessaloniki, GR; [49]Semmelweis University, HU; [50]Tunis Children Hospital, TN; [51]Al Jalila Children's Specialty Hospital, AE; [52]University Medical Center Groningen, NL; [53]University Health Network, CA; [54]University of Toronto, CA, [55]Delft University of Technology, NL.

Introduction

- Progressive familial intrahepatic cholestasis is a rare genetic cholestasis liver disease that affects children. The most common type of PFIC is Bile salt export pump (BSEP) deficiency, which is caused by mutations in the *ABCB11* gene.[1]
- The goal of initial therapy is to relieve pruritus. This is typically achieved using surgical biliary diversion (SBD) or liver transplantation (LTx). [1]
- Markers for the severity of the disease are serum bile acids (sBA) and platelet counts (PLT). However, the longitudinal patterns of sBA and PLT in people with BSEP deficiency are not well characterized.
- Understanding these trajectories is crucial for patient stratification to improve care.

Aim

- Identify longitudinal trajectory patterns of sBA and PLT levels in patients with BSEP deficiency to assess heterogeneity.
- Support the development of targeted therapeutic interventions.
- Enable earlier and more effective treatment by identifying patients at higher risk.

Method

- This study used data from the Natural Course and Prognosis of PFIC and Effect of biliary Diversion (NAPPED) registry, a global cohort comprising 57 sites with longitudinal follow-up of patients with PFIC.[2]
- Children under 18 years were included, while children with benign recurrent intrahepatic cholestasis (BRIC), children with only one mutated allele, and measurements taken after SBD or LTx events were excluded from the analysis.
- To identify subgroups based on sBA and PLT patterns, latent class mixed models were used.[3] These models automatically identify unobserved (latent) subgroups with similar trajectories of longitudinal markers over time. They allow to model individual variation while uncovering meaningful classes in longitudinal data. The sBA values were log-transformed to improve the normality of the residuals.
- The optimal number of latent classes for sBA and PLT trajectories was determined using information criteria and requiring 10% of the study population per subgroup.

NAPPED
Mixed course And Prognosis of PFIC and Effect of biliary Diversion

Table 1 Characteristics of the sBA and PLT trajectory classes. Data are presented as *n* (%) or median (Q1-Q3). BSEP1: at least one allele with p.E297G or p.D482G mutation; BSEP3: 2 predicted protein truncating mutations; BSEP2: all other (missense) mutations.
Abbreviations: SBD, surgical biliary diversion; LTx, liver transplantation; BSEP, bile salt export pump.

Characteristics		Class sBA1	Class sBA2	Class PLT1	Class PLT2
Number of patients		141 (84.4)	26 (15.6)	145 (86.8)	22 (13.2)
Number of patients with first event	SBD	40 (28.4)	2 (7.69)	40 (27.6)	2 (9.09)
	LTx	44 (31.2)	3 (11.5)	37 (25.5)	10 (45.5)
	Death	3 (2.13)	0 (0)	1 (6.90)	2 (9.09)
Sex	Male	61 (43.3)	16 (61.5)	71 (49.0)	6 (27.3)
	Female	78 (55.3)	10 (38.5)	72 (50.0)	16 (72.7)
	Other	1 (0.71)	0 (0)	1 (0.69)	0 (00)
	Unknown	1 (0.71)	0 (0)	1 (0.69)	0 (0)
Follow-up duration, in years		1.25 (0.17-3.71)	6.29 (3.27-13.02)	1.34 (0.17-5.21)	2.67 (1.59-6.83)
BSEP category	BSEP1	42 (29.8)	8 (30.8)	44 (30.3)	6 (27.3)
	BSEP2	62 (44.0)	16 (61.5)	71 (49.0)	7 (31.8)
	BSEP3	37 (26.2)	2 (7.69)	30 (20.7)	9 (40.9)
Number of total measurements		681 (77.9)	193 (22.1)	723 (82.7)	151 (17.3)
Number of measurements per patient		2.00 (1.00-4.00)	4.50 (3.00-9.75)	3.00 (2.00-7.00)	6.00 (4.00-8.75)
Age across total measurements, in years		2.33 (0.75-7.06)	5.82 (2.40-11.17)	4.17 (1.25-8.92)	3.16 (1.25-6.67)

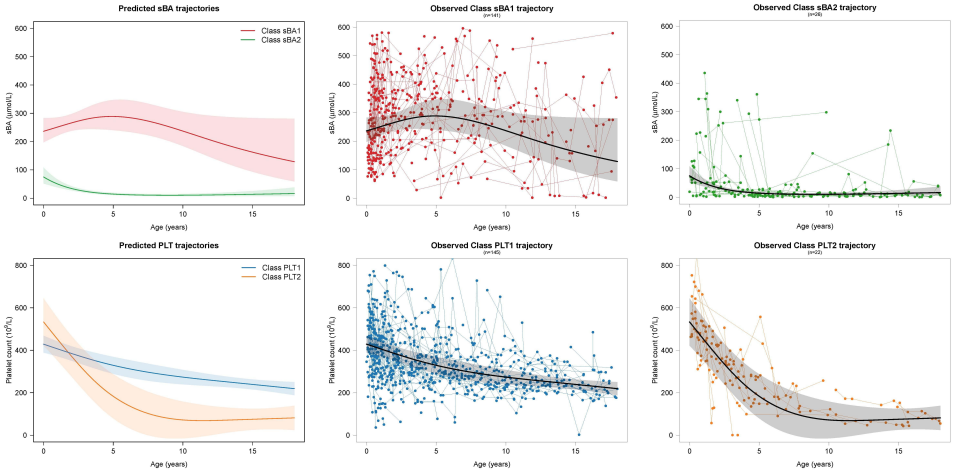


Figure 1 Observed and predicted trajectories of serum bile acid (sBA) and platelet (PLT) counts. Dots represent observed measurements, with observations from the same individuals connected by lines. Bold solid lines and shaded areas indicate the predicted trajectories and their corresponding 95% confidence intervals.

Results

- In both sBA and PLT levels, two mutually exclusive patterns of trajectories were identified (Figure 1).
- Characteristics stratified by class are presented in Table 1.
- An interesting result is Class sBA2, which consists of patients with low sBA values. This class consists mainly of BSEP2 patients (61.5%).
- Only one patient falls into both Class sBA2 and PLT2, highlighting a clear distinction between these two classifications.

Conclusions

- Distinct longitudinal sBA and PLT patterns were identified in patients with BSEP deficiency.
- These patterns reveal substantial heterogeneity in the course of laboratory parameters in time, offering potential for trajectory-specific management strategies that could improve patient care and outcomes.
- Our study identifies a subgroup of BSEP deficiency patterns in the NAPPED registry with (intermittently) relatively low sBA, which requires further analysis.

References

1. Davit-Spraul, A., Gonzales, E., Baussan, C. & Jacquemin, E. (2009). Progressive familial intrahepatic cholestasis. *Orphanet journal of rare diseases, 4*, 1.
2. van Wessel, D. B. E., Gonzales, E., Hansen, B. E., & Verkade, H. J. (2021). Defining the natural history of rare genetic liver diseases: lessons learned from the NAPPED initiative. *European journal of medical genetics, 64(7)*, 104245.
3. Proust-Lima, C., Philipps, V., & Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *Journal of Statistical Software, 78(2)*, 1-56.

Acknowledgements

The authors would like to thank the patients, their families, and their care providers for their contributions to the NAPPED Registry.

Contact:

Figure C.1: Poster about the results of the latent class mixed model presented at the conference EASL